Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterarbeit

Joachim Schole

Strukturierung des Gegenstandsbereichs Whiskysorten mit Hilfe von Textmining

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Joachim Schole

Strukturierung des Gegenstandsbereichs Whiskysorten mit Hilfe von Textmining

Masterarbeit eingereicht im Rahmen der Masterprüfung

im Studiengang Master of Science Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Prof. Dr. Tim Tiedemann

Eingereicht am: 31.08.2018

Joachim Schole

Thema der Arbeit
Strukturierung des Gegenstandsbereichs Whiskysorten mit Hilfe von Textmining

Stichworte
Empfehlungssysteme, Expertensysteme, Knowledge Discovery in Databases, Text Mining, Data Mining, Clustering, Word Embeddings, Whisky

Kurzzusammenfassung
Diese Arbeit befasst sich mit der Herausforderung einer Whisky-Empfehlungsgebung. Es soll die Grundlage für ein fiktives Whisky-Empfehlungssystem geschaffen werden. Konkret ist das Ziel, eine maschinenlesbare Repräsentation von Wissen über geschmackliche Distanzen von Whiskys zu generieren. Zu diesem Zweck wird ein KDD-Prozess festgelegt, welcher auf die Domäne und die verfügbaren Daten zugeschnitten ist. In diesem finden verschiedene Techniken aus dem Text Mining und die Genrerierung von Word Embeddings Anwendung. Ein besonderer Fokus liegt dabei auf der Vorverarbeitung der verwendeten Daten zur Verbesserung der Embeddings. Nach einer Generierung von Repräsentationen von Whiskys, welche Distanzberechnungen erlauben, können Empfehlungen auf Grundlage dieser Distanzen gegeben werden. Diese Empfehlungen werden abschließend im Rahmen der Evaluation durch einige Experten im Hinblick auf ihre Nachvollziehbarkeit bewertet.
Joachim Schole

Title of the thesis
Structuring of the domain whisky varieties with the help of text mining

Keywords
Recommender Systems, Expert Systems, Knowledge Discovery in Databases, Text Mining, Data Mining, Clustering, Word Embeddings, Whisky

Abstract
This thesis addresses the challenge of whisky recommendations. Its goal is to create the basis for a fictional whisky recommendation system. Specifically, the goal of this thesis is to create a machine-readable representation of knowledge about the distances of whiskies regarding their flavour. For this purpose, a KDD process is defined, which is tailored to the domain and the existing raw data. This includes the use of word embeddings. Different techniques from the field

of text mining are applied in the process. A special focus lies on preprocessing the raw text data to improve the resulting word embeddings. After generating representations of whiskies that allow distance calculations, recommendations can be made based on those distances. Finally, these recommendations are assessed in terms of their comprehensibility by several experts as part of the evaluation.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Listings

1. Einleitung

Empfehlungsanfragen stellen für die gefragte Person immer eine Herausforderung dar. Sie muss den Gegenstandsbereich ausreichend gut kennen und möglichst alle in Betracht zu ziehenden Varianten gegeneinander abwägen. Im Fall einer komplexen Domäne wie Whisky ist dies besonders schwierig. Ein gefragter Verkäufer muss gleichzeitig mehrere Faktoren wie Vorrätigkeit, Preisspanne und nicht zuletzt den Geschmack beachten. Letzterer stellt hierbei den komplexesten Faktor dar. Daher findet die Empfehlungsgebung in diesem Bereich oft sehr intuitiv statt. Die aromatische Vielfalt im Bereich Whisky ist groß. Whisky wird heutzutage in vielen Ländern unter Anwendung verschiedener Verfahren hergestellt. Dabei ist ein großes Spektrum verschiedener Ausprägungen entstanden.

Ein Whisky-Verkäufer, welcher beispielsweise nach einem nicht vorrätigen Whisky gefragt wird, muss dem fragenden Kunden sehr spontan eine passende Alternative nennen können. Dies erfordert ein hohes Maß an Erfahrung. Es ist nicht auszuschließen, dass speziell unerfahrenere Verkäufer in einer solchen Situation auf gängige Kategorisierungen zurückgreifen und somit lediglich reproduzieren, was ein Kunde auch selbst recherchieren kann. In diesen und ähnlichen Fällen wäre es hilfreich, ein Software-Tool zur Verfügung zu haben, welches zu einem genannten Whisky eine Auswahl an möglichst ähnlichen Whiskys nennen kann und dabei gegebenenfalls auch überraschende, aber passende Empfehlungen ausspricht. Im Optimalfall kennt ein solches System zudem den Lagerbestand des Geschäftes und nennt ausschließlich vorrätige Whiskys. Um ein solches Expertensystem realisieren zu können, muss zunächst eine maschinenlesbare Repräsentation von Wissen über Whisky vorhanden sein, welche Ähnlichkeiten zwischen einzelnen Whiskys abbildet.

1.1. Ziele und Abgrenzung

Das Ziel dieser Arbeit ist es daher, Möglichkeiten zu überprüfen, ein solches Maschinenwissen zu generieren. Es soll ein Wissensmodell geschaffen werden, welches Distanzberechnungen zwischen Whiskys ermöglicht. Diese Distanzen sollen die geschmackliche Nähe der Whiskys

möglichst realitätsnah abbilden. Es ist ausdrücklich nicht das Ziel dieser Arbeit, ein Produktivsystem zu erstellen. Das konkrete Ziel dieser Arbeit ist es, zu überprüfen, ob mit Machine-Learning-Methoden Ähnlichkeiten zwischen Whiskys ermittelt werden können, welche von Experten akzeptiert werden. Weiter sollen dafür lediglich die geschmacklichen Eigenschaften der Whiskys betrachtet werden. Übergeordnete Daten wie beispielsweise die Herkunft sollen dabei explizit ignoriert werden. Einige der aktuellen Kategorisierungen von Whisky richtigen sich nach Attributen, welche nicht zwangsläufig einen direkten Einfluss auf den Geschmack haben.

1.2. Aufbau der Arbeit

Zunächst findet eine Recherche zu Empfehlungs- und Expertensystemen statt. Daraufhin folgt eine Betrachtung von Möglichkeiten zur Wissensgewinnung für Expertensysteme. Um die Zwischenergebnisse dieser Arbeit besser bewerten zu können, erfolgt im Anschluss eine eingehende Recherche der Domäne Whisky. Diese bildet bereits den ersten Schritt eines für diese Arbeit erstellen KDD-Prozesses. Dieser wird im Rahmen der Arbeit mehrmals durchlaufen, um nach Betrachtungen der jeweiligen Zwischenergebnisse gegebenenfalls Anpassungen an vorigen Schritten vorzunehmen. Dabei finden verschiedene gängige Methoden aus dem Text Mining Anwendung. Am Ende des Wissensgewinnungsprozesses steht ein Set aus maschinenlesbaren Repräsentationen von Whiskys als Ergebnis. Dieses ermöglicht Distanzberechnungen zwischen einzelnen Whiskys. Auf Basis dieser können Empfehlungen generiert und Klassen gebildet werden. Um eine Aussage über die Qualität dieser Empfehlungen treffen zu können, erfolgt anschließend eine Befragung von Experten zur Nachvollziehbarkeit dieser. Nach dieser Evaluation erfolgt eine Analyse weiterer Optimierungsschritte und ein abschließendes Fazit zu dieser Arbeit.

2. Analyse

Die Entwicklung von Expertensystemen bietet vor allem dort einen Mehrwert, wo Expertenwissen nicht oder nur begrenzt verfügbar ist. Zudem bieten Expertensysteme den Vorteil, dass sie im Gegensatz zu menschlichen Experten konsistente Entscheidungen treffen (Cooke, 1994). Besonders durch das Aufkommen der Forschung im Bereich der Künstlichen Intelligenz hat der Prozess der Abbildung von Wissen und Intelligenz in Maschinen an Bedeutung gewonnen (Cooke, 1994). Diese Arbeit beschreibt die Entwicklung einer Wissensgrundlage für ein Empfehlungssystem. Empfehlungssysteme sind eine Form von Expertensystemen. Für die Entwicklung einer Wissensgrundlage für Expertensysteme findet üblicherweise eine *Knowledge Acquisition* (Wissensgewinnung) statt. Den zentralen Teil der Knowledge Acquisition bildet die *Knowledge Elicitation* (Wissenserhebung). Diese kann je nach Anwendungsfall aus verschiedenen Methoden bestehen.

2.1. Empfehlungssysteme

Ziel dieser Arbeit ist die Schaffung einer Wissensgrundlage für ein hypothetisches Empfehlungssystem. Dabei liegt der Fokus auf der Entwicklung des maschinellen Wissens als Grundlage für ein solches System. Weiter soll dieses Wissen explizit nicht durch manuell entwickelte Wissensmodelle, sondern durch maschinelles Lernen erfolgen. Eine Einführung in Empfehlungssysteme bietet Ricci u. a. (2011). Aus dieser Quelle stammen die wesentlichen Aussagen in diesem Kapitel.

Empfehlungssysteme gewinnen durch die große Auswahl an Produkten im Internet immer mehr an Bedeutung und sind durch ihre Anwendung in diversen beliebten Online-Diensten bereits allgegenwärtig. Schon die Suche nach einem schlichten Alltagsprodukt kann einen Nutzer überfordern, da er teilweise hunderte von Alternativen präsentiert bekommt. Ein durchschnittlicher Nutzer hat nicht die Zeit oder Muße, sich bei einem Einkauf im Internet mit einer derartigen Menge an Möglichkeiten auseinanderzusetzen. An diesem Punkt setzen Empfehlungssysteme an. Empfehlungssysteme sind Software-Lösungen, welche zum Zweck haben, einem Nutzer nützliche oder interessante Objekte vorzuschlagen. Der Begriff des Objektes kann

dabei beliebig abstrakt interpretiert werden. So kann ein Empfehlungssystem beispielsweise auch Spielzüge empfehlen und somit die Grundlage einer künstlichen Intelligenz für ein Computerspiel bilden. Andere Beispiele für mögliche Empfehlungsobjekte sind Nachrichtenartikel oder Musikstücke. In dieser Arbeit bilden Whiskys die Objekte. Es gilt, ein Ähnlichkeitsmaß für Whiskys zu entwickeln, um anhand eines genannten Whiskys ähnliche Whiskys berechnen zu können. Ein möglicher Anwendungsfall für ein Whisky-Empfehlungssystem ist bereits in Kapitel 1 beschrieben: Ein Kunde, welcher in einem Spirituosengeschäft nach einem Whisky fragt, der sich nicht auf Lager befindet. Der Verkäufer könnte in diesem Fall auf ein Whisky-Empfehlungssystem zurückgreifen, welches ihm die geschmacklich nächstgelegenen Whiskys anzeigt. Aus diesen kann der Verkäufer dem Kunden Vorschläge nennen, ohne selbst über Expertenwissen im Bereich Whisky verfügen zu müssen. Im Optimalfall wäre das System so eingerichtet, dass es nur Whiskys für die Empfehlung in Betracht zieht, welche das Geschäft auf Lager hat. Menschen, welche keine oder wenig Erfahrung in einer Domäne haben, bilden die Hauptzielgruppe von Empfehlungssystemen.

Sowohl für den Nutzer als auch den Betreiber bietet ein Empfehlungssystem einen Mehrwert. Für gut befundene Empfehlungen befriedigen den Kunden und erhöhen dadurch die Chance, dass er dem Betreiber treu bleibt. Allerdings lässt sich diese Annahme nur bedingt auf den oben genannten Beispielanwendungsfall übertragen. Gerade in solchen Fällen kann der direkte Austausch mit dem Verkäufer zu einer höheren Kundenbindung führen, sofern der Verkäufer sich als guter Empfehlungsgeber und damit als Experte beweist.

Es existieren verschiedene mögliche Varianten, auf welcher Datengrundlage Empfehlungssysteme nützliche Objekte ermitteln. Je nach Domäne sind einige dieser Varianten besser geeignet als die anderen. Ein sehr verbreiteter und von der Domäne relativ unabhängiger Ansatz ist das kollaborative Filtern, welches dem Nutzer Objekte empfiehlt, die anderen Nutzern mit ähnlichem Geschmack bereits gefallen haben. Dieser Ansatz eignet sich besonders für Systeme, welche beispielsweise Medien wie Musik und Filme empfehlen. Generell ist das kollaborative Filtern eine beliebte Methode für Empfehlungen in Online-Shops. Einen ähnlichen Ansatz verfolgen gemeinschaftsbasierte Systeme. Solche Systeme betrachten anstatt von Nutzern mit ähnlichem Geschmack befreundete Nutzer und deren Vorlieben. Diesem Ansatz liegt die Annahme zugrunde, dass die Nutzer ihren Freunden bei einer Empfehlung eher vertrauen als fremden Nutzern. Der Nachteil dieses Ansatzes ist es, dass solche Empfehlungssysteme die soziale Vernetzung des Nutzers in irgendeiner Form kennen müssen. In bestimmten Domänen kann es sinnvoll sein, gewisse Eigenschaften des Nutzers wie sein alter und seinen Wohnort zu

betrachten. Solche Systeme heißen demographische Empfehlungssysteme. Die Beachtung des Alters des Kunden verhindert, dass ihm für seine Altersgruppe ungeeignete Objekte empfohlen werden. Inhaltsbasierte Empfehlungssysteme sind Systeme, welche die Eigenschaften der Objekte miteinander vergleichen und dem Nutzer solche Objekte empfehlen, welche anderen Objekten ähneln, die der Nutzer in der Vergangenheit für gut befunden hat. Alternativ kann ein solches System eine Empfehlung auch anhand eines vom Nutzer genannten Referenzobjekts oder anhand von Wunscheigenschaften geben. Die komplexeste Variante stellen wissensbasierte Empfehlungssysteme dar. Diese Systeme basieren auf einem tiefen Wissen über die Domäne, um Objekte nach einem Problemlösungsprinzip oder durch eine automatische Bedarfsermittlung zu empfehlen. Der Nachteil dieses Ansatzes ist ein sehr hoher Pflegebedarf und gegebenenfalls die Notwendigkeit der Kenntnis einiger relevanter Eigenschaften des Nutzers. Um einige der jeweiligen Nachteile der verschiedenen Ansätze zu verringern besteht die Möglichkeit, hybride Empfehlungssysteme zu entwickeln. Diese Systeme bilden Kombinationen aus beliebig vielen der genannten Varianten.

Beispiele für bestehende Whisky-Empfehlungssysteme sind Distiller (2018), Whisky.de (2018) und Whiskyology (2018). Eine eingehendere Betrachtung dieser Systeme findet in Kapitel 2.3.6 statt.

2.2. Wissensgewinnung für Expertensysteme

In der Regel finden zur Bildung von Experten- beziehungsweise Wissenssystemen Techniken aus dem Bereich der Knowledge Elicitation Anwendung. Eine Einführung und Kategorisierung dieser Techniken bietet Cooke (1994). Aus dieser Quelle stammen die grundlegenden Informationen in Kapitel 2.2.1. Eine Alternative zur Knowledge Elicitation ist die automatisierte Wissensermittlung. Eine Methode dafür ist die *Knowledge Discovery in Databases* nach Fayyad u. a. (1996a) und Fayyad u. a. (1996b). Diese ist in Kapitel 2.2.2 beschrieben.

2.2.1. Knowledge Elicitation

Der Prozess der Knowledge Elicitation bildet den wesentlichen Teil der Knowledge Acquisition. Die Knowledge Elicitation beschreibt dabei die initiale Datenerhebung, für die verschiedene Methoden angewandt werden können. Die Knowledge Acquisition beschreibt den generellen Prozess der Abbildung von Wissen in Formaten, welche anschließend von Maschinen verwendet werden können. Den Methoden der Knowledge Elicitation ist gemein, dass sie alle eine Form

des Dialogs zwischen Experte und *Elicitor* - im Folgenden Ermittler genannt - darstellen. Das aus dem Prozess resultierende Wissensmodell kann insofern als Vermittler zwischen Experte und Ermittler betrachtet werden, als dass es das Wissen in einem Format darstellt, das beide verstehen und über dessen Korrektheit sich beide einig sind.

Aufgrund des teilweise sehr hohen Aufwands und der gleichzeitig großen Bedeutung der Wissenserlangung wird diese auch als Flaschenhals in der Entwicklung eines Expertensystems bezeichnet. Eine besondere Schwierigkeit ergibt sich aus der Tatsache, dass Experten oftmals intuitiv handeln und daher nicht oder nur schwer artikulieren können, aus welchen Grund sie gerade welche Entscheidung treffen. Zudem neigen Menschen allgemein dazu, beim vermitteln von Wissen dieses zu vereinfachen, um es dem Gegenüber verständlicher zu machen. Weiterhin besteht die Gefahr, dass ein Experte verschiedenes Wissen in der Wertigkeit für seine Arbeit falsch einschätzt und somit weniger wichtige Aspekte hervorhebt und wichtigere unbeabsichtigt unterschlägt.

Knowledge-Elicitation-Techniken lassen sich in die drei Familien *Beobachtung / Interviews*, *Prozessverfolgung* und *Konzeption* aufteilen. Generell lässt sich dabei sagen, dass verschiedene Techniken je nach Anwendungsfall besser beziehungsweise schlechter geeignet sind.

Beobachtung / Interviews Die Techniken aus der ersten Familie stellen sehr direkte Vorgehensweisen dar. Der Experte wird entweder direkt in einem Interview befragt oder aber bei der Durchführung verschiedener Aufgaben beobachtet. Die Varianten der Interviews reichen dabei von komplett offen bis hin zu sehr starr strukturierten Befragungen. Bei der Beobachtung des Experten bieten sich ebenso verschiedene Varianten an. Diese richten sich vor allem nach der Art der Aufgabe, welche von alltäglichen bis hin zu absoluten Ausnahme- und Notfällen reichen kann. Je nach Aufgabenstellung kann der Ermittler aktiv an der Durchführung der Aufgabe teilnehmen oder aber den Experten passiv beobachten. Hauptkritikpunkt an diesen Varianten ist der potentielle Einfluss der Anwesenheit des Ermittlers. Weiterhin besteht die Möglichkeit einer Aufgabenanalyse, bei der Aufgaben in Zusammenarbeit mit dem Experten in Unteraufgaben aufgeteilt werden, um den sequentiellen Ablauf einer Aufgabe abzubilden.

Allgemein lässt sich sagen, dass die Techniken aus dem Bereich Beobachtung / Interviews besonders für die frühe Phase der Knowledge Elicitation geeignet sind. Ein großer Nachteil ist, dass die Interpretation der Ergebnisse dieser Techniken sehr schwierig sein kann. Dem gegenüber steht der vergleichsweise sehr geringe Vorbereitungsaufwand.

Prozessverfolgung Die Techniken aus dieser Familie stellen Vorgehensweisen dar, welche einen starken Fokus auf die Durchführung einzelner Aufgaben legen. Sie unterscheiden sich von den beobachtenden Aufgaben insofern, dass die Form der aufgezeichneten Daten vorab definiert ist. Der Fokus auf spezielle Aufgaben beinhaltet unter anderem das Risiko, dass die für die Erhebung ausgewählten Aufgaben nicht repräsentativ für die alltägliche Arbeit des Experten sind sondern tendenziell eher Spezialfälle abdecken. Die Varianten der Techniken dieser Familie umfassen verbale und non-verbale Berichte, Protokollanalysen und Entscheidungsanalysen. Die verbalen Berichte unterscheiden sich weiter in solche, welche während der Aufgabendurchführung und solche, welche im Anschluss an die durchgeführte Aufgabe aufgezeichnet werden. Beide Varianten beinhalten das Risiko, dass der Experte ihm offensichtlich erscheinende Informationen auslässt. Zudem ist eine Berichterstattung zeitgleich zur Bearbeitung nicht bei allen Aufgaben möglich. Dagegen ist eine nachträgliche Berichterstattung stark vom Gedächtnis des Experten abhängig. Non-verbale Berichte umfassen beispielsweise das *Eye-Tracking* während der Bearbeitung. Die Entscheidungsanalyse bezeichnet die Ermittlung der Entscheidungsfindung des Experten. Die zentralen Daten dabei sind die Faktoren, welche die Entscheidung beeinflussen und ihre jeweilige Gewichtung.

Konzeption Die Techniken aus dieser Familie verfolgen das Ziel, Wissen und Konzepte aus der betrachteten Domäne strukturell und relational darzustellen. Eine Hauptgruppe in dieser Familie bildet die Konzepterhebung. Bei diesen Methoden fällt der Hauptteil der konzeptionellen Arbeit dem Experten zu. Dieser kann beispielsweise darum gebeten werden, wesentliche Konzepte aus seiner Domäne aufzulisten oder ein Inhaltsverzeichnis für ein fiktives Buch über seine Domäne zu verfassen. Weitere Methoden lassen den Experten die Verhältnisse zwischen verschiedenen Konzepten in seiner Domäne bewerten oder aber hierarchische Strukturen aufbauen.

Generell sind die Methoden aus dieser Familie besonders gut dafür geeignet, Daten von mehreren Experten zu sammeln und zusammenzutragen. Einen Nachteil kann die sehr abstrakte Betrachtungsweise der Arbeit des Experten bilden, mit der dieser gegebenenfalls Schwierigkeiten haben könnte.

Diese Aufteilung zeigt nur einen Teil der möglichen Methoden zur Wissenserhebung. Generell unterscheiden sich die meisten Methoden in ihrem Grad der Formalität. Methoden, welche zu formlos oder zu formell sind, bringen am ehesten Nachteile mit sich. Eine alternative Möglichkeit

der Aufteilung ist die, die Methoden zunächst in analytische und synthetische Methoden zu unterteilen (Boose, 1989). Dies würde in der oben beschriebenen Struktur die ersten beiden Familien zusammenfassen.

Eignung der Knowledge Elicitation

Im Bezug auf die Domäne Whisky stellen sich einige Schwierigkeiten bei einer Möglichen Anwendung der erläuterten Methoden heraus. Die Hauptaufgabe des Systems stellt die Empfehlung eines Whiskys dar. Da dies ein gedanklicher und vor allem sehr intuitiver Prozess ist, ist über eine reine Beobachtung des Experten kaum nützliches Wissen ermittelbar. Auch eine Befragung eines Experten in Interviews erscheint nicht vielversprechend. Da eine Empfehlung auch keinem artikulierbaren Muster folgt, sind die Methoden aus der Prozessverfolgung ebenfalls ungeeignet.

Deutlich geeigneter erscheinen die konzeptionellen Methoden. Ein durchaus geeigneter Ansatz wäre es, eine Wissensarchitektur in Form von Kategorisierungen und Verhältnissen von Whiskys zueinander aufzubauen. Der größte Nachteil dieser Methode liegt vor allem im großen Aufwand, welcher der Aufbau einer solchen Struktur mit sich bringt. Eine besondere Schwierigkeit liegt außerdem in der Verfügbarkeit von Experten für die Wissenserhebung. Eine Alternative zur Befragung von Experten bildet die Verwendung von Dokumenten (Cooke, 1994). Diese sind im Bereich der Domäne Whisky vorhanden. Hier stellt sich allerdings die Schwierigkeit, das Wissen aus diesen Dokumenten zu extrahieren und daraus ein intelligentes System zu schaffen. Dies kann mit Methoden des Machine Learnings geschehen, welches dann die Knowledge Acquisition im Aufbauprozess des Empfehlungssystems weitestgehend ersetzt. Allerdings ist auch hier die Einbindung von Expertenwissen zur Validierung der Ergebnisse notwendig. Es stellt also einen Prototyp-basierten Ansatz dar (Cooke, 1994). Eine Schablone für einen solchen Prozess bildet der allgemeine Prozess der *Knowledge Discovery in Databases*.

2.2.2. Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) ist eine Prozessschablone, die von Fayyad u. a. (1996a) konzipiert wurde. Sie gibt eine grobe Abfolge von Schritten vor, durch die Wissen aus (Roh-) Datenmengen bezogen werden kann, welche für eine manuelle Auswertung zu groß sind. Nach jedem Schritt des Prozesses besteht die Möglichkeit zur Rückkehr zu einem beliebigen vorigen Schritt, um Anpassungen vorzunehmen. Abbildung 2.1 zeigt den Aufbau eines KDD-Prozesses nach Fayyad u. a. (1996a). Die wesentlichen Aussagen in diesem Kapitel stammen, wenn nicht anders vermerkt, neben Fayyad u. a. (1996a) aus Fayyad u. a. (1996b) und Sharafi (2013).

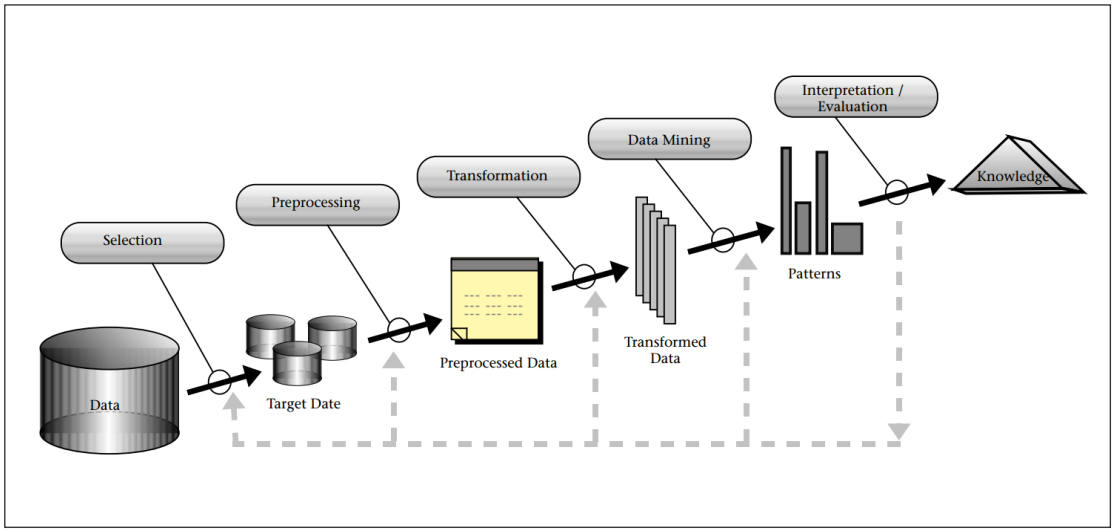


Abbildung 2.1.: Der KDD-Prozess (Fayyad u. a., 1996a, S. 41)

Vor Beginn des eigentlichen KDD-Prozesses muss ein ausreichend tiefes Wissen über die betrachtete Domäne vorhanden sein, um das Ziel des Prozesses und das genaue Vorgehen ausführlich definieren zu können. Außerdem müssen mögliche Probleme und Risiken bekannt sein. Die Erlangung dieses Wissens kann je nach Domäne einen Großteil der gesamten Arbeitszeit ausmachen. Eine Einführung in die Domäne Whisky und die Beschaffenheit der zugrundeliegenden Daten bietet Kapitel 2.3.

Unter Anwendung des erlangten domänenspezifischen Wissens erfolgt im ersten Schritt eine Recherche und Bewertung verschiedener möglicher Datenquellen. Auf Grundlage der so erfolgten Bewertung wird ein Teil der verfügbaren Daten zur Verwendung im Prozess ausgewählt. Dabei sind verschiedene Faktoren wie Qualität, Relevanz und Repräsentativität der Daten ausschlaggebend. Kapitel 3.2.1 beschreibt Recherche, Vergleich und Auswahl verschiedener Datenquellen aus der Domäne Whisky. Je nach weiterem Vorgehen ist es sinnvoll, das Datenset in Trainings- und Testdaten zu unterteilen.

Der Schritt der Vorverarbeitung (*Preprocessing*) umfasst je nach Beschaffenheit der Daten und geplanter weiterer Vorgehensweise verschiedene Methoden zur Aufbereitung der Daten. Im Falle von Textdaten beinhaltet dies in der Regel das Entfernen von *Stopwords* und störender Interpunktion sowie gegebenenfalls die Reduktion sämtlicher Wörter auf ihren Stamm. Eine Betrachtung einiger für diese Arbeit eventuell relevanten Methoden bietet Kapitel 2.4. Des

Weiteren sind die verwendeten Methoden in den Kapiteln 3.3 bis 3.6 an den entsprechenden Stellen technisch näher erläutert.

Die Transformation beschreibt die Umwandlung der aufbereiteten Daten in für die weitere Verarbeitung geeignete Repräsentationsmodelle. Dies kann ebenfalls durch verschiedene Methoden geschehen, welche unterschiedliche Formen von Repräsentationsmodellen zum Ergebnis haben können. Zu den Möglichen Repräsentationsmodellen gehören unter anderem einfache Zahlenwerte, Vektoren beliebiger Dimensionalität und andere, komplexere Modelle wie beispielsweise Baumstrukturen. Die in dieser Arbeit angewandten Methoden zur Transformation sind in den Kapiteln 2.6 und 3.9 beschrieben.

Auf dem so entstandenen Datenset aus Repräsentationsmodellen können in der Folge im Zuge des Data Minings verschiedene Algorithmen angewandt werden. Das Data Mining ist ein zentraler Bestandteil des KDD-Prozesses. Es dient der Erkennung von Mustern im aufbereiteten Datenset. Je nach Repräsentationsform und Fragestellung existieren verschiedene Algorithmen, welche angewandt werden können. Ziel des Data Minings ist entweder die Entdeckung neuen Wissens oder die Verifikation aufgestellter Hypothesen. Data-Mining-Algorithmen unterteilen sich in die Kategorien Klassenbildung *Clustering*, Klassifizierung *Classification*, Assoziationsanalyse und Zeitreihenanalyse. Clustering und Klassifizierung unterscheiden sich insofern, dass das Clustering automatisch Klassen in den Daten ermittelt, während die Klassifizierung lediglich eine Zuordnung zu bestehenden Klassen durchführt. Eine Assoziationsanalyse ermittelt häufig gemeinsam auftretende Daten. Diesen Algorithmen gemein ist, dass sie zur Erkennung von Mustern in den Daten dienen. Ein solches Muster kann beispielsweise eine Wahrscheinlichkeit für ein Ereignis oder eine Aufteilung der Daten in Gruppen sein, welche gegebenenfalls eine neue Kategorisierung darstellen oder eine alte bestätigen. Durch die Interpretation und Evaluation dieser Muster kann neues Wissen erlangt werden.

Clustering

Ein Ziel dieser Arbeit ist es, anhand der in ihre Repräsentationsform übertragenen Whiskys neue Kategorien und Klassen zu ermitteln. Die hierfür angewandte Methode ist das Clustering, welches die Data-Mining-Methode im KDD-Prozess darstellt. Im Allgemeinen bestehen Clustering-Prozesse aus den Phasen Festlegung der Repräsentationsform (*Pattern Representation*), Definition der Distanzfunktion (*Pattern Proximity*), *Clustering*, Datenabstraktion (*Data Abstraction*) und der Bewertung der Ergebnisse (*Assesment of Output*) (Jain u. a., 1999). Diese sind im Folgenden näher beschrieben.

Festlegung der Repräsentationsform Vor dem Clustering muss eine Definition der Repräsentationsform der zu verarbeitenden Daten stattfinden. In der Regel erwarten Clustering-Algorithmen Daten in Form von *Feature Vektoren*. Die Elemente dieses Vektors können quantitative, qualitative, ordinale Werte oder auch strukturelle Werte wie Bäume sein. In der Regel erwarten Clustering-Algorithmen einfache Vektoren, welche einen Datensatz in einem multidimensionalen Raum verorten.

Definition der Distanzfunktion Ebenfalls vor dem Clustering muss eine Distanzfunktion definiert werden, welche dem Algorithmus vorgibt, wie die Feature Vektoren im Verhältnis zueinander stehen. Im Fall von komplexeren Daten muss an dieser Stelle gegebenenfalls eine eigene Distanzfunktion entwickelt werden. Dies bietet auch die Möglichkeit, einzelne Features selbst zu gewichten. Im Regelfall stellen die Feature Vektoren eine Projektion der Datenobjekte in einen multidimensionalen Raum dar. Liegen die Feature Vektoren in dieser Form vor, ist es ausreichend, ein Vektordistanzmaß auszuwählen. Häufig verwendete Distanzmaße sind die euklidische und die Kosinus-Distanz. Die euklidische Distanz beschreibt den Abstand zweier Punkte im Raum, während die Kosinus-Distanz den Kosinus des Winkels θ zwischen zwei Vektoren beschreibt.

$$d_2(a,b) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2} \tag{2.1}$$

$$\cos(\theta) = \frac{a \cdot b}{|| \ a \ ||_2 \cdot || \ b \ ||_2} = \frac{\sum_{i=1}^{d} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{d}(a_i)^2} \cdot \sqrt{\sum_{i=1}^{d}(b_i)^2}} \tag{2.2}$$

Gleichung 2.1 zeigt die euklidische Distanz der Vektoren a und b mit der Dimensionalität d. Gleichung 2.2 zeigt die Kosinus-Distanz derselben (Kosinus-Ähnlichkeit, 2018). Um zu verhindern, dass einige Features andere dominieren, da ihre Werte über größere Skalen verteilt sind, können die Feature Vektoren vor der Anwendung der euklidischen Distanz normalisiert werden. Eine Normalisierung der Feature Vektoren führt zu einer Annäherung der euklidischen an die Kosinus-Distanz.

Clustering In der eigentlichen Clustering-Phase findet die Anwendung des Cluster-Algorithmus auf die in den vorigen Schritten vorbereiteten Daten unter Verwendung der definierten Distanzfunktion statt. Clustering Algorithmen teilen sich je nach Methode in verschiedene Kategorien auf. Harte Clustering-Methoden ordnen jeden Datensatz einem Cluster zu.

Unscharfe Methoden errechnen dagegen das Maß der Zugehörigkeit jedes Datenobjekts zu jedem Cluster. Weiter stehen hierarchische Methoden partitionierenden entgegen. Erstere erstellen einen hierarchischen Cluster-Baum, in dem die Datenobjekte die Blätter bilden. Partitionierende Methoden teilen die Objekte auf einer Ebene in Gruppen auf. Hierarchische Methoden können entweder agglomerativ oder zerteilend arbeiten. Agglomerative Algorithmen beginnen mit einem Cluster pro Datenobjekt und fügen diese nach und nach zusammen. Zerteilende Algorithmen beginnen mit einem einzigen Cluster, in dem alle Datenobjekte liegen und zerteilen diesen Schritt für Schritt. Zudem wird zwischen monothetischen und polythetischen Verfahren unterschieden. Erstere teilen Datenobjekte anhand des Vergleichs einzelner Features auf. Letztere vergleichen immer die gesamten Feature Vektoren der Objekte. Distanzfunktionen wie die euklidische und die Kosinus-Distanz eignen sich nur für polythetische Methoden, da diese die Distanz von Lokalisierungen in einem Raum repräsentieren.

Datenabstraktion Die Datenabstraktion ist ein optionaler Schritt des Clusterings. Abhängig von der weiteren Vorgehensweise müssen die Daten in eine dafür geeignete Form gebracht werden. Zu den möglichen weiteren Schritten gehören die Bewertung der Cluster durch Menschen oder durch Bewertungsmetriken. Die meisten Bewertungsmetriken erwarten eine *Ground Truth* zum Vergleich der Cluster mit einer realen Kategorisierung oder einem anderen Clustering-Ergebnis. Um die Bewertung durch Menschen zu ermöglichen muss eine geeignete Abstraktionsform der Cluster festgelegt werden. Diese kann beispielsweise die Auswahl repräsentativer Datensätze sein. Dabei können entweder der Mittelpunkt eines Clusters oder auch am Rande des Clusters gelegene Datensätze verwendet werden. Eine weitere Möglichkeit ist die Ermittlung der Häufigkeit bestimmter Eigenschaften der Datenobjekte eines Clusters.

Bewertung der Ergebnisse Die Bewertung der Ergebnisse ist ebenfalls ein optionaler Schritt. Eine Möglichkeit bieten die bereits erwähnten Bewertungsmetriken zum Vergleich mit bestehenden Kategorien. Eine Übereinstimmung der Cluster mit Bestehenden Kategorien bietet zwar keine neuen Erkenntnisse, bestätigt aber die Richtigkeit des vorangehenden Prozesses. Eine alternative Möglichkeit ist die Bewertung der Cluster durch Experten. Dies schließt die Gewinnung neuen Wissens nicht in dem Maße aus wie der Vergleich mit bestehenden Kategorien. Gleichzeitig ist die Bewertung von Experten immer auch subjektiven Empfindungen ausgesetzt. Dieser Nachteil lässt sich teilweise durch die Befragung möglichst vieler Experten eindämmen. Des Weiteren besteht bei einer Expertenbefragung

immer das Risiko, dass diese nicht in vollem Umfang verstehen, was die konkrete Fragestellung ist. Um dies zu verhindern müssen die Anforderungen an die Experten klar formuliert werden.

2.3. Die Domäne Whisky

Um sowohl in Frage kommende Datenquellen als auch die Ergebnisse dieser Arbeit ausreichend bewerten zu können, ist es wichtig, ein eingehendes Verständnis der Domäne Whisky zu erlangen. Dabei sind besonders die bestehenden Kategorien und die aromatische Zusammensetzung von Whiskys von Interesse. Die bestehenden Kategorien dienen als Vergleichsmöglichkeit für die Bewertung der Ergebnisse dieser Arbeit. Die Betrachtung der aromatischen Zusammensetzung dient der Möglichkeit, die Ergebnisse auf einer tieferen Ebene als den bestehenden Kategorien zu bewerten.

2.3.1. Entstehung

Die erste Erwähnung von Whisky als *Aqua Vitae* befindet sich in den Aufzeichnungen des obersten schottischen Schatzmeisters aus dem Jahr 1494, wobei der Begriff Aqua Vitae zu dieser Zeit für die meisten alkoholischen Destillate gebräuchlich war. Nach damaligem Glauben besaßen diese Getränke heilende Wirkungen. Die lateinische Bezeichnung hat sich für den Aquavit bis heute gehalten. Der Begriff Whisky geht auf die gälische Übersetzung *uisge beatha* zurück (Bathgate, 2003).

Zunächst lag das Recht, Brände herzustellen ausschließlich bei den schottischen Klöstern. Nach der Auflösung dieser durch Henry VIII begannen die meisten größeren Häuser und Höfe, sich ihre eigenen Brennkessel zur Herstellung von Alkohol für verschiedene Zwecke zu beschaffen. Die daraufhin entstehende Whisky-Industrie erlebte durch die industrielle Revolution zunächst einen Aufschwung. Durch Ernteausfälle folgten Brennverbote und Abwanderungen von Brennern nach Kanada und in die Vereinigten Staaten. Aufgrund der agrikulturellen Bedingungen in diesen Ländern erfolgte dort ein Wechsel von Gerste zu Roggen als Rohmaterial, was zur Entstehung von Bourbon und Rye Whiskey führte. Zu Beginn galt Whisky als Spirituose der ärmeren Schichten. Aufgrund mehrerer schlechter Weinernten in Folge in Frankreich und der damit einhergehenden Knappheit von Brandy rückte Whisky auch in den Fokus der wohlhabenderen Bevölkerung. Dies löste ein starkes weltweites Wachstum in der Whisky-Industrie aus, welches in einem Kollaps endete, den die Industrie überlebte (Stewart u. a., 2014).

Heutzutage ist schottischer Whisky durch den Scotch Whisky Act (1988), die Scotch Whisky Regulations (2009) und durch EU-Recht (Regulation (EC) No 110/2008, 2008) streng reguliert und geschützt. Die wichtigsten Voraussetzungen für einen Whisky, um als Scotch bezeichnet werden zu dürfen, sind die ausschließliche Verwendung von Wasser und Getreide, eine minimale Reifedauer in einem Eichenfass von drei Jahren, ein Mindestalkoholgehalt von 40% und die Produktion in Schottland.

2.3.2. Bestehende Kategorien

Es existieren mehrere Eigenschaften, anhand derer sich Whiskys kategorisieren lassen. Die populärste Unterscheidung ist die nach der Herkunft des Whiskys. Hier wird im Allgemeinen nach schottischem (Scotch) Whisky, irischem Whiskey und amerikanischem Whiskey unterschieden, wobei amerikanischer Whiskey in dieser Aufteilung oft synonym zu Bourbon gebräuchlich ist. Whiskys aus anderen Ländern bilden dabei eine eigene Kategorie. Diese Aufteilung hat ihren Ursprung in der Geschichte des Whiskys, welche in Kapitel 2.3.1 beschrieben ist. Diese oberflächliche Aufteilung nach der Herkunft bildet allerdings die Vielfalt von Whiskys nicht ausreichend ab. Der schottische Whisky teilt sich weiter anhand der Regionen auf, in denen er produziert wird. Diese Regionen sind *Islay*, *Highlands*, *Lowlands*, *Campbeltown*, *Speyside*. Eine weitere Kategorie bilden die Inseln ohne Islay. Amerikanischer Whisky teilt sich weiter in kanadischen Whisky, Bourbon, Kentucky Whiskey und Tennessee Whiskey auf. Dies sind dabei lediglich die wichtigsten Kategorien. Eine weitere, häufig gesondert genannte Produktionsregion ist Japan. Andere Länder werden in der Regel als „Rest der Welt“ zusammengefasst, wobei es in vielen Ländern eine wachsende Whisky-Industrie gibt (Stewart u. a., 2014).

Eine weitere, näher am Geschmack orientierte Kategorisierung ist die Unterteilung von Whiskys nach der verwendeten Getreideart. Innerhalb Schottlands wird dabei lediglich zwischen *Malt* und *Grain* Whiskys unterschieden. Dabei werden Malt Whiskys aus gemalzter Gerste und Grain Whiskys aus ungemalzten anderen Getreidesorten - in der Regel Weizen - hergestellt (Bringhurst und Brosnan, 2014). Der Zusatz *Single* steht für Whiskys, die vollständig in einer einzigen Destillerie hergestellt wurden (Scotch Whisky Regulations, 2009). Blend Whiskys entstehen durch die Vermischung mehrerer Whiskys und bilden den Hauptanteil an verkauftem Whisky. Blends stehen für einen immer konstanten Geschmack während ungemischte Whiskys sich je nach Abfüllung unterscheiden. Weitere häufig verwendete Getreidesorten sind Roggen für Rye Whiskey und Mais in unterschiedlichen Anteilen für Bourbon und Corn Whiskey.

Generell wird für die meisten Whisky Sorten eine Mischung aus mindestens zwei Getreidearten verwendet.

Andere Attribute, anhand derer Whiskys in Kategorien unterteilt werden können sind die Dauer der Fassreife, der verwendete Fasstyp und der Alkoholgehalt. Diese Attribute haben einen direkten Einfluss auf das Aroma und den Geschmack des Whiskys. Eine Aufteilung nach diesen Eigenschaften ist allerdings nur bedingt sinnvoll, da sie sich gegenseitig bedingen. Zudem ist die Dauer der Fassreife immer in Kombination mit der Herstellungsregion zu betrachten, da diese in wärmeren Gebieten beschleunigt stattfindet (Jackson, 2017).

2.3.3. Herstellungsprozess

Whisky zeichnet sich durch eine besonders komplexe aromatische und geschmackliche Zusammensetzung aus. Diese erlangt er in einem langwierigen Herstellungsprozess, welcher unter anderem in Jack (2014) beschrieben ist. Aus dieser Quelle stammen die wesentlichen Aussagen dieses Kapitels.

Whisky-Hersteller müssen sich auf dem Markt voneinander abgrenzen und sich durch eine eigene Identität auszeichnen. Kunden erwarten in der Regel schon durch den Namen der Destillerie oder des Abfüllers eine bestimmte Geschmacksrichtung. Jeder einzelne Schritt in der Herstellung eines Whiskys bietet verschiedene Möglichkeiten, den Geschmack des Endproduktes zu beeinflussen. Die Whisky-Herstellung teilt sich grob in die Schritte Gärung, Destillation und Fassreife auf. Dabei hat bereits die Auswahl der Rohstoffe einen entscheidenden Einfluss auf die Qualität und aromatische Zusammensetzung des Endproduktes. Als Rohmaterial dienen immer Getreide und Wasser. Aus dem Getreide bezieht der Whisky Aromen wie *Malz* und *Biskuit*. Das Wasser wurde traditionell aus einem naheliegenden Fluss bezogen. Daher war zu Beginn der Whisky-Herstellung schon die Wahl des Standorts der Destillerie entscheidend. Dies ist heutzutage nicht mehr entscheidend, da ausreichend alternative Wasserquellen zur Verfügung stehen. Dolan (2003) und Bringhurst u. a. (2014) beschreiben die zu beachtenden Aspekte bei der Auswahl einer Wasserquelle.

Ein weiterer Standortaspekt war die Beschaffenheit des lokalen Torfs, sofern dieser verwendet wurde. Torf dient als Brennmaterial während des Trocknens des Malzes, um dieses lagerbar zu machen. Je nach Abbauregion des Torfes besteht dieser aus den teilweise zersetzten Resten verschiedener Pflanzen. Der verwendete Torf hat einen direkten Einfluss darauf, welche Aromen während des Torfens in das Malz übergehen. Diese Aromen sind *Torfrauch*, *Phenol*, *Rauch*, *verbrannt* und *medizinisch* (Bringhurst und Brosnan, 2014).

Während der Gärung entstehen verschiedene Ester, Aldehyde und Säuren. Diese variieren je nach Hefestamm, Dauer der Gärung und der Temperatur. Ester sind vor allem als fruchtige Aromen erkennbar. Aldehyde finden unter anderem in der Parfumherstellung Verwendung. Während der anschließenden Destillation verfliegen je nach Siedepunkt einige Aromen wieder. Außerdem finden unter anderem Maillard-Reaktionen statt, welche auch während des Bratens von Speisen auftreten. Diese lassen beispielsweise Aromen wie *Getreide*, *Blumen*, *Gras*, *verbrannt* und *schweflig* entstehen. Den letzten Herstellungsschritt bildet die Fassreife. In dieser Zeit bezieht der Whisky neben diversen Aromastoffen auch Farbe aus dem Fass. Je nach Holzart und vorheriger Nutzung des Fasses bezieht der Whisky in unterschiedlichen Verhältnissen vorwiegend Vanillin und Eichenlaktone - auch Whisky-Laktone genannt. Letztere bringen Aromen wie *Kokosnuss*, *Erde*, *Holz*, *Heu* und *Sellerie*. Einige Flüchtige Aromen verfliegen während der Fassreife mit dem sogenannten *Angel's Share*. Ein weiterer aus dem Fass bezogener Stoff ist Eugenol, welches ein Gewürznelkenaroma hat. Tannine werden ebenfalls bezogen. Diese bewirken ein trockenes Mundgefühl, auch Adstringenz genannt. Die Auswahl des richtigen Fasstyps ist von großer Bedeutung für das Endprodukt. In der Regel finden Fässer aus europäischer oder amerikanischer Eiche Verwendung. Dabei unterscheiden sich die Fässer weiter in ihrer Vorverarbeitung. Conner (2014) unterscheidet zwischen den Fasstypen *New Charred*, *Ex-Sherry*, *Ex-Bourbon*, *Refill* und *Regenerated*. *Charring* bezeichnet das anbrennen der inneren Schicht des Fasses. Zur Regenerierung wird die verkohlte Schicht eines benutzten Fasses weitestgehend entfernt und die darunterliegende Schicht erneut gebrannt. Dies bringt allerdings Einbußen bei der Geschmacksentwicklung mit sich. In Ex-Sherry-Fässern wurde zuvor ein Sherry gereift, in Ex-Bourbon-Fässern ein Bourbon-Whiskey. Je nach Eichenart und Vornutzung geben die Fässer unterschiedliche Aromen ab (Conner, 2014).

Zusätzlich zu den beschriebenen Herstellungsschritten erlaubt die Regulierung die Zugabe von Zuckercouleur, um die Farbe des Whiskys anzupassen. Nach jedem der Beschriebenen Herstellungsschritte wird eine Qualitätskontrolle durchgeführt. Diese erfolgt aufgrund des hohen Alkoholgehalts allein durch riechen. Zudem suggeriert Piggott und Jardine (1979), dass eine Unterscheidung von Whiskys allein anhand der Gerüche ausreichend genau ist. Das sogenannte Nosing findet ebenfalls beim Blending Anwendung.

2.3.4. Geschmackliche Zusammensetzung

Aus den oben beschriebenen Herstellungsschritten ergibt sich ein sehr komplexes Geschmacksbild für Whisky. Zusätzlich zu den genannten Aromen werden auch die geschmacklichen

Hauptmerkmale *süß*, *bitter*, *sauer* und das Mundgefühl betrachtet. Neben adstringent kann ein Whisky auch ein wärmendes oder vollmundiges Mundgefühl haben. Adstringenz ist ein Begriff aus der Weinsprache und bezeichnet ein raues, pelziges Mundgefühl. Ein weiterer Faktor, welcher die Wahrnehmung eines Whiskys beeinflusst, ist seine Farbe (Piggott und Jardine, 1979).

Abbildung 2.2.: Das Nosing Wheel des Scotch Whisky Research Institute (Jack, 2014, S. 238)

Die Komplexität der geschmacklichen Zusammensetzung von Whisky ist anhand von Nosing Wheels erkennbar. Diese wurden als Versuch hergestellt, ein einheitliches Vokabular für Tasting Notes zu schaffen. Außerdem zeigen Nosing Wheels eine hierarchische Aufteilung der Aromen. Abbildung 2.2 zeigt ein Beispiel für ein Nosing Wheel. Die Aromen sind in 12 Hauptkategorien unterteilt. Dazu bilden die Primären Geschmäcker, das Mundgefühl und nasale Effekte jeweils

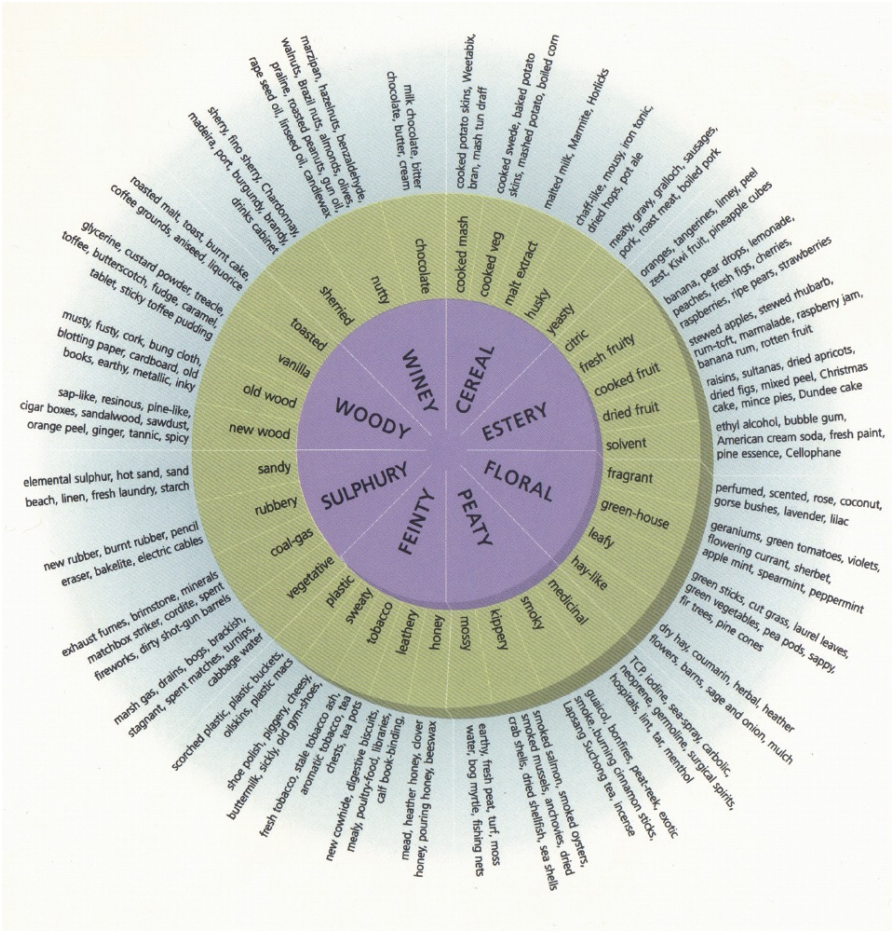


Abbildung 2.3.: Ein weiteres Nosing Wheel nach Charles MacLean (Wishart, 2009, S. 24)

eine eigene Kategorie. Der Aufbau des Nosing Wheels legt nahe, dass dieses sich als Grundlage für eine Geschmacksontologie eignet. Eine solche Ontologie bildet eine Möglichkeit zur Ermittlung von Distanzen zwischen Geschmacksbegriffen. Allerdings setzt dies voraus, dass das Vokabular des Nosing Wheels die in den Tasting Notes verwendeten Geschmacksbegriffe vollständig beinhaltet. Dies erscheint anhand der Anzahl der Begriffe im Nosing Wheel nicht realistisch. Die Hinzunahme weiterer Nosing Wheels aus anderen Quellen schafft weitere Probleme. Die Struktur von Nosing Wheels ist nicht einheitlich, was die Vereinigung mehrerer Exemplare zu einem erschwert und sehr intuitiv macht. Hierfür ist Expertise im Bereich der Lebensmittelforschung erforderlich. Abbildung 2.3 zeigt ein weiteres Nosing Wheel. Dieses

verfügt über weniger Hauptkategorien. Die Hauptgeschmacksrichtungen, das Mundgefühl und die nasalen Effekte sind ausgelassen. Zudem sind einige Unterschiede in den Zuordnungen erkennbar. Im ersten Nosing Wheel befindet sich der Begriff *cooked vegetables* auf unterster Ebene in der Kategorie *sulfury*, während der gleiche Begriff im zweiten Nosing Wheel eine Unterkategorie der Hauptgeschmacksrichtung *cereal* bildet. Des Weiteren enthält das erste Nosing Wheel eine Kategorie *cask derived*. Dies bildet keine Strukturierung nach Geschmacksrichtungen ab. Generell lässt sich sagen, dass die Struktur von Nosing Wheels nicht immer nachvollziehbar erscheint. Eine Ontologie, die sich hiernach richtet, würde die tatsächliche Nähe von Geschmacksbegriffen nicht ausreichend abbilden. Daher ist die Verwendung einer auf Nosing Wheels basierenden Ontologie nicht geeignet.

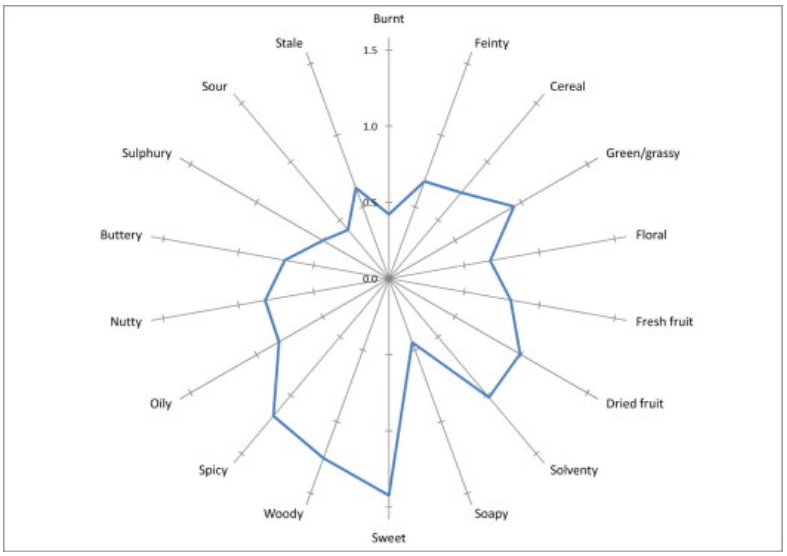


Abbildung 2.4.: Ein Flavour Profile (Jack, 2014, S. 241)

2.3.5. Beschreibungsformen

Die gängigste Form, Whiskys zu beschreiben, sind Tasting Notes. Dies sind formlose Fließtexte, welche je nach Autor nur aus einer Aneinanderreihung von Aromen oder aus einem längeren Text mit zusätzlichen Informationen bestehen. Die erste Form ist dabei verbreiteter. Im weitesten Sinne lassen sich auch die Herstellerbeschreibungen auf Whisky-Verpackungen als Tasting Notes bezeichnen. Diese sind allerdings immer zu Werbezwecken verfasst und daher in ihrer Bewertung beeinflusst. Eine alternative Beschreibungsform sind *Flavour Profiles*. Dies sind

quantifizierte Beschreibungen von Whiskys. Ein Beispiel für ein Flavour Profile zeigt Abbildung 2.4.

Ein Nachteil der Flavour Profiles ist offensichtlich die Beschränkung auf eine limitierte Anzahl an Aromen. Dies macht Flavour Profiles einerseits vager als Tasting Notes. Dem gegenüber steht allerdings die sehr klare Quantifizierung dieser Aromen, was wiederum in Tasting Notes deutlich vager ist. Hier stehen nur quantifizierende Begriffe zur Verfügung, was deutlich subjektiver und schwer in Zahlen zu übersetzen ist. Dennoch birgt die Beschränkung des Vokabulars auf wenige Begriffe die Gefahr, einige Eigenschaften über- oder unterzubewerten. Das Flavour Profile in Abbildung 2.4 zeigt 18 verschiedene Begriffe, Piggott und Jardine (1979) nennt allerdings nach einer eingehenden Analyse eine Liste von 35 Begriffen. Dennoch beschreibt Jack und Steele (2002), dass ein Neuronales Netz, welches auf Flavour Profiles mit lediglich 13 Geschmacksrichtungen trainiert wurde, besser zwischen Scotch und anderen Whisky-Sorten unterscheidet als menschliche Experten.

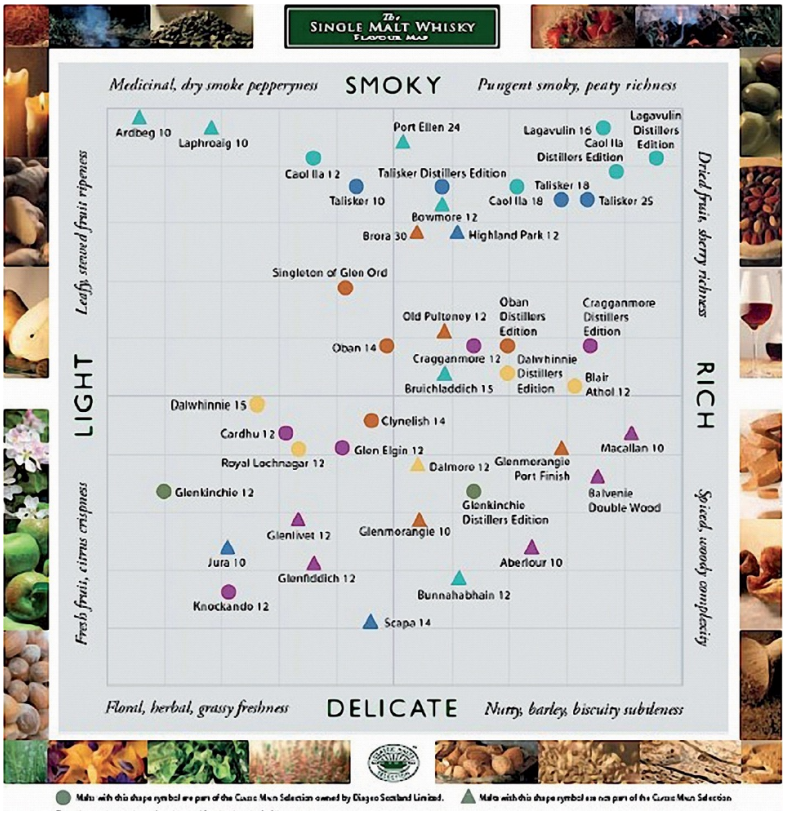


Abbildung 2.5.: Eine Single Malt Whisky Flavour Map nach Diageo (Wishart, 2009, S. 25)

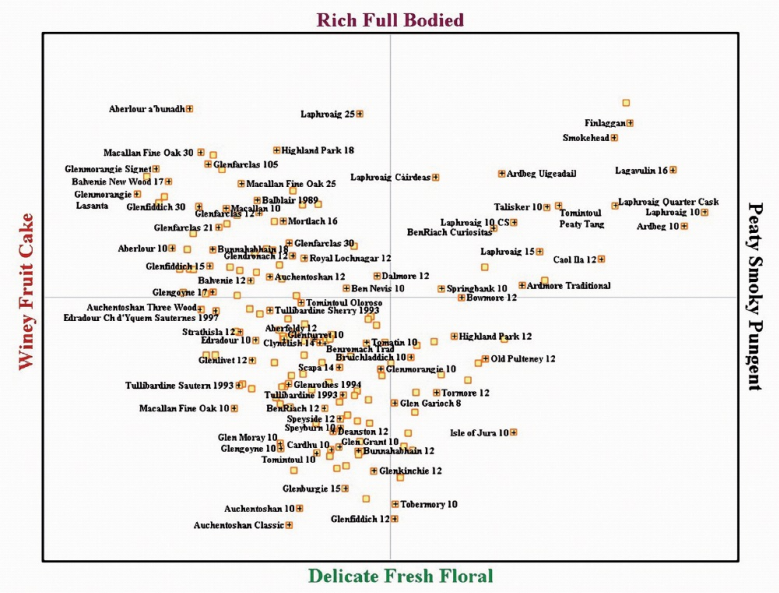


Abbildung 2.6.: Eine Flavour Map als Ergebnis einer Analyse der Hauptaromen von Whiskys (Wishart, 2009, S. 26)

Als weitere Beschreibungsform von Whisky können sogenannte Flavour Maps bezeichnet werden. Dies sind Grafiken, in denen die Nähe von Whiskys untereinander dargestellt werden soll. Gleichzeitig ordnen Flavour Maps Whiskys in einer Art Geschmacksraum an. Dabei stellen sie allerdings jeweils sehr allgemeine Geschmacksrichtungen gegenüber. Die Abbildungen 2.5 und 2.6 zeigen Beispiele für solche Flavour Maps. Dabei stellt letztere das Ergebnis einer Datenanalyse dar. Diese erfolgte auf Basis von quantitativen Einschätzungen von Hauptgeschmacksrichtungen im Bereich Whisky (Wishart, 2009).

2.3.6. Bestehende Whisky-Empfehlungssysteme

Es existieren bereits einige Empfehlungssysteme für Whisky, die verschiedene Ansätze verfolgen. Die Empfehlungsanwendung Distiller (2018) bietet ihren Nutzern die Möglichkeit, sich in einem Webbrowser oder per mobiler Anwendung zu einer Whisky-Empfehlung zu navigieren. Dabei fragt das System den Nutzer zunächst, ob er eher einen neuen, unbekannteren Whisky oder einen etablierten Whisky sucht. Daraufhin fragt das System nach dem Kenntnisstand des Nutzers und nach dem Anlass und der Umgebung, in der der Kunde vorhat, den Whisky zu trinken. Das System bietet dabei auch andere Arten von Spirituosen an. Während eine Unterteilung nach „Einsteiger“- und „Fortgeschrittenen“-Whiskys durchaus nachvollziehbar ist, erscheint die

Unterscheidung nach Anlass und Umgebung eher subjektiv. Zudem erfragt das System auch die Region des gewünschten Whiskys, was ein Mindestmaß an Vorwissen beim Nutzer voraussetzt.

Whiskyology (2018) verfolgt einen ähnlichen Ansatz, bietet aber nur Whiskys an. Das System fragt den Kunden zunächst, nach welchem Kriterium er die Whiskys suchen möchte. Daraufhin folgt eine tiefere Befragung, die von dem ausgewählten Kriterium abhängt. Das System verfügt dabei lediglich über 50 verschiedene Whiskys.

Beiden genannten Systemen ist gemein, dass sie Whiskys nach anerkannten, aber auch nach sehr subjektiven Kriterien unterteilen. Zudem fragen beide Systeme den Kunden nach der gewünschten Preisspanne. Der Aufbau und der Betrieb dieser Systeme erfordert ein hohes Maß an Datenpflege durch die Anbieter. Jedem Whisky müssen weitere übergeordnete Kriterien hinzugefügt werden, was viel Wissen und Intuition erfordert. Die genauen Prozesse hinter diesen Systemen sind selbstverständlich nicht einsehbar.

Einen anderen Ansatz verfolgt der deutsche Anbieter Whisky.de (2018). Dieser bietet dem Anwender die Möglichkeit, sich ein Geschmacksprofil zusammenzustellen und damit die Whiskys in der Datenbank des Systems zu filtern. Die Geschmacksprofile der Whiskys werden dabei aus Bewertungen von Nutzern errechnet. Dieses System setzt allerdings ein gewisses Maß an Vorkenntnis beim Nutzer voraus.

Alle drei genannten Systeme sind entweder stark von der Pflege der Daten durch den Anbieter oder von den Vorkenntnissen des Nutzers abhängig.

Einen dieser Arbeit ähnlichen Ansatz verfolgt Krzus (2018). Dieses System verwendet ebenfalls Word Embeddings zur Generierung der Empfehlungen. Die wesentlichen Unterschiede zwischen dem System und dieser Arbeit liegen im verwendeten Datenset und in der Vorverarbeitung. Zudem folgt diese Arbeit dem KDD-Prozess.

2.4. Betrachtung der Rohdaten

Im Rahmen des für diese Arbeit konzipierten KDD-Prozesses wurden während der Recherche verschiedene Mögliche Datenquellen erfasst. Dieses Kapitel betrachtet die Beschaffenheit der Daten und insbesondere die der bezogenen Tasting Notes. Da als Datenquelle Online-Ressourcen ausgewählt wurden, liegen die Daten zunächst als HTML-Dokumente vor. Aus diesen werden die Tasting Notes und die verfügbaren Metadaten bezogen. Je nach Ressource liegen dabei unterschiedliche Metadaten vor.

In der Regel bestehen Tasting Notes fast nur aus einer Aneinanderreihung von Geschmacksbegriffen. In einigen Fällen bestehen die Tasting Notes allerdings auch aus ganzen Fließtexten,

welche noch zusätzliche Informationen enthalten. Diese Informationen sind in dieser Arbeit nicht weiter von Interesse. Es sollen lediglich die in der Tasting Note benannten geschmacklichen Eigenschaften des beschriebenen Whiskys ermittelt werden. Oft werden Whiskys in Tasting Notes auch mit anderen Whiskys verglichen. Drei unterschiedliche Beispiele für zufällig aus dem Datenset ausgewählte Tasting Notes sind im Folgenden gezeigt.

> *„Nose: Perfumy. Liquorice, white pepper. On time, luscious sweet notes swirl up. Vanilla, pear drop. Quite complex and finely laced.*
> *Palate: Silky. But firm. The palate is not as enticing as the nose. A pleasant fruitiness at start then a bitter oaky flavour.*
> *Finish: Medium, slightly spiced.*
> *Comment: The nose was promising but the palate did not deliver all. Water stresses the bitterness.“*
>
> *Benromach 8 Years Old Maderia* (Whisky Magazine, 2018)

> *„Nose: Broad beans, mint and a hint of olive, which in turn has some added Brazil nut. Firm but mature. In time, things become drier with hay/raffia and balsa wood blandness. When water's added, there's a surprising whiff of Sharpie pen and peanut butter.*
> *Palate: Initially, this shows good maturity with a red fruit element beginning to peek out, before that Tormore crunch comes in and locks such frivolity away.*
> *Finish: Rigid.*
> *Conclusion: It was all going well, but Tormore's a tough customer.“*
>
> *Tormore 28 Years Old* (Scotchwhisky.com, 2018)

> *„This whisky is very much like eating a candy apple next to a fireplace. Fruity and vanilla notes are further enriched by rock candy sweetness. Light baking spices, orange peel and toffee are also present, then a quick plumage of sweet burning wood wafts in. This would be an ideal Scotch to sip with dessert or to finish a meal.“*
>
> *Arran 18 Years Old* (Distiller, 2018)

Es lassen sich Unterschiede in der Strukturierung der Tasting Notes erkennen. Während die ersten beiden Beispiele wie oft üblich nach *Nose*, *Palate* und *Finish* aufgeteilt sind, besteht das letzte Beispiel aus einem Fließtext ohne Unterscheidung zwischen Gerüchen und Mundgefühl. Für diese Arbeit ist das Ziel, die Texte so weit zu reduzieren, dass sie möglichst nur noch aus

Geschmacksbegriffen in ihrer Grundform bestehen. Grammatikalische Eigenschaften der Texte sind zunächst nicht weiter interessant. Folgend sind die Tasting Notes noch einmal in dieser Wunschform gezeigt.

> *„perfume, liquorice, white pepper, luscious, sweet, vanilla, pear drop, complex, silky, firm, fruit, bitter, oak, spice“*

> *„broad beans, mint, olive, brazil nut, firm, mature, dry, hay, raffia, balsa wood, bland, sharpie pen, peanut butter, mature, red fruit, crunch, rigid“*

> *„candy apple, fireplace, fruit, vanilla, rock candy, sweet, baking spices, orange peel, toffee, sweet, burning wood“*

An diesen Texten ist bereits erkennbar, dass ein Ähnlichkeitsmaß für Geschmacksbegriffe unbedingt notwendig ist, um die Vergleichbarkeit zu verbessern. Ein einfacher Vergleich der vorkommenden Begriffe führt hier lediglich zu dem Schluss, dass die oberen beiden Whiskys in der Eigenschaft Härte (*firm*) übereinstimmen. Ein angewandtes Ähnlichkeitsmaß zwischen den Begriffen führt dazu, dass beispielsweise die Begriffe *fruit* und *red fruit* als ähnlich bewertet werden und somit auch hier ein weiterer Faktor für die Ähnlichkeit der beiden Whiskys erkannt wird. Zudem verbessert ein solches Ähnlichkeitsmaß auch die Genauigkeit, mit der die Ähnlichkeit der beiden Texte ermittelt wird. Generell lässt sich sagen, dass anhand der gezeigten Texte keine große Ähnlichkeit zwischen den beiden oberen Whiskys erkennbar ist. Zwischen dem ersten und dem dritten Whisky sind deutlich mehr Übereinstimmungen erkennbar.

2.5. Einfluss verschiedener Vorverarbeitungsschritte auf Texte

Ein Teilziel dieser Arbeit ist es, Tasting Notes in eine einheitliche Form zu bringen. Hierfür müssen Methoden erprobt werden, welche zu einer Annäherung an die in Kapitel 2.4 vorgeschlagene Wunschform führen.

Entfernung von Stoppwörtern

Eine weit verbreitete Methode zur Textvorverarbeitung ist das Entfernen von Stoppwörtern. Dies ist dann sinnvoll, wenn durch die Stoppwörter kein Mehrwert an Informationen zu erwarten ist. Im englischen betrifft dies unter anderem Wörter wie *a*, *the*, und *and*. Die drei Tasting Notes aus Kapitel 2.4 sind im Folgenden nach Anwendung eines Stoppwort-Filters dargestellt. Davor

fand eine Entfernung der Satzzeichen und eine Konversion aller Buchstaben in Kleinbuchstaben statt. Die Kommentare fanden keine weitere Berücksichtigung.

> *„perfumy liquorice white pepper. time luscious sweet notes swirl. vanilla pear drop. quite complex finely laced silky firm palate enticing nose. pleasant fruitiness start bitter oaky flavour medium slightly spiced.“*

> *„broad beans mint hint olive turn added brazil nut. firm mature. time things become drier hay raffia balsa wood blandness. water's added there's surprising whiff sharpie pen peanut butter. initially shows good maturity red fruit element beginning peek tormore crunch comes locks frivolity away. rigid.“*

> *„whisky much like eating candy apple next fireplace. fruity vanilla notes enriched rock candy sweetness. light baking spices orange peel toffee also present quick plumage sweet burning wood wafts. would ideal scotch sip dessert finish meal.“*

Die Beispiele zeigen, dass bereits eine starke Annäherung der Texte an die Wunschform erreicht ist. Dennoch ist an einigen Stellen Verbesserungspotential erkennbar.

Stemming

Anhand der Beispieltexte nach Stoppwort-Entfernung ist erkennbar, dass eine Verallgemeinerung der Wortformen notwendig ist, um Wörter wie *mature* und *maturity* als gleich zu erkennen. Eine Möglichkeit, dies zu erreichen ist das *Stemming*. Stemming bezeichnet die Rückführung von Wörtern auf ihren Stamm. Folgend stehen die Beispieltexte nach Anwendung des Stemming.

> *„perfumi liquor white pepper. time lusciou sweet note swirl. vanilla pear drop. quit complex fine lace silki firm palat entic nose. pleasant fruiti start bitter oaki flavour medium slightli spice.“*

> *„broad bean mint hint oliv turn ad brazil nut. firm matur. time thing becom drier hay raffia balsa wood bland. water' ad there' surpri whiff sharpi pen peanut butter. initi show good matur red fruit element begin peek tormor crunch come lock frivol away. rigid.“*

> *„whiski much like eat candi appl next fireplac. fruiti vanilla note enrich rock candi sweet. light bake spice orang peel toff also present quick plumag sweet burn wood waft. would ideal scotch sip dessert finish meal.“*

Wie an den Beispieltexten erkennbar ist, führt das Stemming dazu, dass die beiden Wörter *mature* und *maturity* nun in der gleichen Schreibweise vorliegen und somit in der weiteren Verarbeitung auch als gleich erkannt werden. Dieser Schritt ist in dieser Arbeit möglich, da die grammatikalische Form der Wörter nicht von Relevanz ist. In anderen Kontexten kann diese sehr wichtig sein, weshalb das Stemming nicht immer ein geeignetes Mittel ist. Eine Schwäche des Stemmings ist am zweiten Beispieltext erkennbar. Apostrophe werden nicht entfernt. Weiterhin ist erkennbar, dass einige Begriffe nicht gänzlich auf ihren Stamm zurückgeführt werden. Der Begriff *oaki* würde nicht als gleich mit dem Begriff *oak* erkannt werden.

Phrasenerkennung

An den Beispieltexten ist weiterhin erkennbar, dass viele Begriffe, welche sich aus mehreren Wörtern zusammensetzen, nicht als zusammengehörig erkannt werden. Ein Mittel, um zusammengehörende Wörter zu ermitteln ist die Erkennung von Phrasen. Diese ermittelt je nach verwendetem Algorithmus zu jedem Wortpaar im Datensatz die Häufigkeit, in der die Wörter in Kombination auftreten und markiert solche als Paar, bei denen dies oft der Fall ist. Begriffe, die aus mehr als zwei Wörtern bestehen, lassen sich durch mehrfache Anwendung dieses Algorithmus ermitteln. Im Folgenden sind die drei Beispieltexte nach mehrfach angewandter Phrasenerkennung gezeigt.

> *„perfumi liquor white_pepper. time lusciou sweet note swirl. vanilla pear_drop. quit complex fine lace silki firm palat entic nose. pleasant fruiti start bitter oaki flavour medium slightli spice.“*

> *„broad_bean mint hint oliv turn ad brazil_nut. firm matur. time thing becom drier_hay_raffia_balsa wood bland. water' ad there' surpri_whiff_sharpi_pen_peanut_butter. initi show good matur red fruit element begin peek_tormor_crunch come lock_frivol_away. rigid.“*

> *„whiski much like eat candi appl next fireplac. fruiti vanilla note enrich_rock candi sweet. light bake_spice orang_peel toff also present quick_plumag sweet burn wood waft. would ideal scotch sip dessert finish meal.“*

Die Beispieltexte zeigen, dass Begriffe wie *white_pepper*, *pear_drop* und *brazil_nut* korrekt als zusammengehörig erkannt werden. Allerdings ist ebenso erkennbar, dass Begriffe wie *red fruit* und *candi appl* nicht erkannt werden, dafür aber *drier_hay_raffia_balsa*. In letzterem Fall

wäre die Zuordnung von *balsa* und *wood* zueinander richtig. Dies lässt sich vermutlich darauf zurückführen, dass die zusammengefassten Wörter jeweils nur sehr selten im gesamten Textkorpus vorkommen, wohingegen der Begriff *wood* deutlich häufiger auch in anderen Kontexten vorkommt.

2.6. Ermittlung von Distanzen zwischen Datensätzen

Die errechnete Distanz zweier Whiskys soll nach Fragestellung dieser Arbeit die Distanz ihrer Geschmäcker möglichst genau darstellen. Da als Rohdaten Tasting Notes verwendet werden, müssen diese in eine Repräsentationsform transformiert werden, die Distanzberechnungen ermöglicht. Eine einfache, naheliegende Möglichkeit wäre es, die Übereinstimmung zweier Tasting Notes anhand der gemeinsam auftretenden Wörter zu errechnen. Diese Methode ist allerdings nicht ausreichend, da sie die Ähnlichkeit verschiedener Geschmacksbegriffe zueinander nicht abbildet. So müssen Begriffspaare wie *fruit*, *berries* und *spice*, *pepper* beispielsweise in sich näher beieinander liegen als beliebige andere Paare aus den Begriffen. Anders formuliert muss die Distanz zwischen *fruit* und *berries* geringer sein als die zwischen *fruit* und *spice*. Eine Möglichkeit, diese Ähnlichkeiten abzubilden, sind Ontologien. Diese sind für den Gegenstandsbereich Whisky nicht verfügbar und die Erstellung einer eigenen Ontologie für diese Arbeit ist nicht realisierbar. Eine weitere Möglichkeit bildet die Generierung von Word Embeddings. Dies beschreibt die Berechnung von Vektorrepräsentationen zu Wörtern. In dem dadurch entstehenden multidimensionalen Raum liegen ähnliche Wörter nahe beieinander. Wortvektoren bilden somit eine potentielle Methode, Distanzen zwischen Geschmacksbegriffen zu ermitteln.

Als Algorithmen zur Berechnung von Word Embeddings existieren *Word2Vec*, *FastText* und *GloVe* (Mikolov u. a., 2013b; Bojanowski u. a., 2016; Pennington u. a., 2014). In dieser Arbeit findet Word2Vec Anwendung. Dieser Algorithmus basiert auf den Algorithmen *Skip-Gram* und *Continuous Bag-of-Words* (CBOW) (Mikolov u. a., 2013a). Ein gemeinsamer Nachteil der genannten Algorithmen ist die Notwendigkeit eines ausreichend großen Trainingsdatensets. Im Folgenden ist die Funktionsweise des Word2Vec-Algorithmus näher erläutert. Die wesentlichen Aussagen stammen - sofern nicht anders angegeben - aus Mikolov u. a. (2013a) und Mikolov u. a. (2013b).

Der Word2Vec-Algorithmus trainiert ein Neuronales Netz darauf, zu jedem Wortpaar in einem Textkorpus die Wahrscheinlichkeit zu ermitteln, im selben Kontext zu stehen. Kontext beschreibt dabei den Bereich vor und nach einem Wort in einem Satz. Zur Berechnung der Wahrscheinlichkeiten bietet der Algorithmus die erwähnten Varianten Skip-Gram und CBOW

an. Erstere errechnet zu einem Wortpaar (a, b) die Wahrscheinlichkeit, mit der ausgehend von der Präsenz eines Wortes a davon auszugehen ist, dass ein zufällig aus dem Kontext des Wortes a ausgewähltes Wort das Wort b ist. Die zweite Variante dagegen errechnet aus einem Kontext k die Wahrscheinlichkeit für die Präsenz eines Wortes w. Durch das Training des Netzes entsteht eine interne Gewichtungsmatrix. Aus dieser Matrix können die Embeddings der Wörter des Textkorpus entnommen werden.

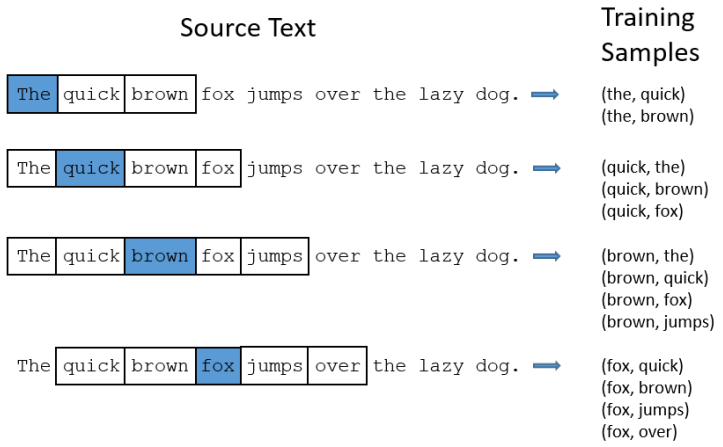


Abbildung 2.7.: Kontextbasierte Wortpaarbildung (McCormick, 2016)

Initial muss der Textkorpus so aufbereitet werden, dass zu jedem Wort Paare mit allen in seinem Kontext vorkommenden Wörtern gebildet werden. Das eigentliche Training geschieht also auf einem Set aus Wortpaaren. Die Größe des Kontextes ist dabei variabel. Abbildung 2.7 zeigt ein Beispiel für eine Wortpaarbildung mit einer Kontextgröße von 2. Ein größerer Kontext verringert den Einfluss von Füllwörtern, welche tendenziell keinen oder nur einen geringen Informationsgehalt haben. Die Größe der zu ermittelnden Wortvektoren ist ebenfalls setzbar. Diese ist identisch mit der Anzahl der Neuronen im Hidden Layer des Netzes. Dieser Prozess ermittelt zu Wörtern, welche Häufig in sehr ähnlichen Kontexten verwendet werden, sehr ähnliche Wortvektoren. Umgekehrt kann also aus nahe beieinanderliegenden Vektoren darauf geschlossen werden, dass die von ihnen repräsentierten Wörter häufig in sehr ähnlichen Kontexten vorkommen. Ausgehend davon, dass häufig in ähnlichen Kontexten verwendete Wörter eine ähnliche Bedeutung haben, stellt die Entfernung zweier Wortvektoren voneinander eine semantische Distanz der entsprechenden Wörter dar.

Ein Beispiel dafür sind die beiden Sätze *the bunny jumps* und *the kangaroo jumps*. Die beiden Wörter *bunny* und *kangaroo* erscheinen im gleichen Kontext. Dies spricht, ohne die Bedeutung der einzelnen Wörter kennen zu müssen, dafür, dass beide Wörter etwas gemeinsam haben. In diesem Fall beschreiben Beide Begriffe ein Tier, welches in der Lage ist, zu springen. Unter der Annahme, dass ähnliche Whiskys von unterschiedlichen Experten ähnlich beschrieben werden, kann dieses Prinzip auch auf Tasting Notes übertragen werden. Beschreibt ein Tester einen Whisky mit den Geschmacksbegriffen *peat, wood, berries, vanilla* und ein zweiter mit *vanilla, peat, fruit, wood*, so kann davon ausgegangen werden, dass die Worte *berries* und *fruit* eine gewisse semantische Ähnlichkeit haben.

Die in diesem Kapitel beschriebene Logik ist auch auf andere Datentypen anwendbar. McCormick (2018) fasst einige Beispiele zusammen, wie Unternehmen unter anderem Musikhörverläufe von Nutzern mit Hilfe des beschriebenen Prinzips analysieren und darauf basierend Empfehlungen generieren (Karam, 2017). Der Word2Vec-Algorithmus lässt sich also auch als Methode zur Analyse zeitlicher Abläufe verwenden. Dadurch bilden die resultierenden Embeddings auch Wahrscheinlichkeiten für die Vorhersage zukünftigen Hörverhaltens.

3. Experimente

Dieses Kapitel beschreibt die im Rahmen dieser Arbeit durchgeführten Experimente. Der Fokus liegt dabei auf der Optimierung der Word Embeddings. Die Durchführung der Experimente orientiert sich am zuvor erstellten KDD-Prozess. Dieser beinhaltet auch Rückschritte zu vorigen Stufen innerhalb des Prozesses. Der KDD-Prozess ist Produkt vorangegangener Arbeiten (Schole, 2017a,b, 2018). In Schole (2018) ist der Prozess beschrieben und evaluiert. Es hat bereits ein kompletter Durchlauf des Prozesses stattgefunden. Zu Beginn dieser Arbeit soll der Prozess unter Berücksichtigung der im ersten Durchlauf erlangten Erkenntnisse verfeinert und erneut durchlaufen werden. Im Folgenden Kapitel ist der Prozess anhand der einzelnen Schritte aus Kapitel 2.2.2 beschrieben.

3.1. Aufbau eines auf die Domäne und das Arbeitsziel zugeschnittenen KDD-Prozesses

Vor Beginn der Arbeit und der Bestimmung des KDD-Prozesses muss ein ausreichendes Wissen über die Domäne vorhanden sein. Eine Recherche zur Domäne Whisky hat bereits in vorigen Arbeiten stattgefunden (Schole, 2017b). Das dabei erlangte Wissen legt die Grundlage für eine tiefere Recherche im Rahmen dieser Arbeit. Das Ergebnis einer erneuten, tieferen Recherche ist in Kapitel 2.3 festgehalten.

Der anschließende Schritt ist die Auswahl der zu verwendeten Rohdaten. Bevor diese erfolgen kann, muss zunächst eine umfassende Recherche möglicher Quellen erfolgen. Nach einem Vergleich zwischen verschiedenen Büchern (Murray, 2016; Ronde, 2016) und Online-Quellen fällt die Wahl auf die Verwendung letzterer. Der große Vorteil der Online-Quellen ist ihre Verfügbarkeit und vor allem die Menge und Vielfalt der vorhandenen Daten. In vorigen Arbeiten fand ein genauerer Vergleich dieser Möglichkeiten statt (Schole, 2017b). Die Recherche und Auswahl der in dieser Arbeit verwendeten Datenquellen ist in Kapitel 3.2.1 tiefer beschrieben.

Zur den möglichen Schritten zur Vorverarbeitung der Daten gehören verschiedene Methoden aus dem *Natural Language Processing*. Darunter fallen unter anderem die gängigen Prozesse

Stemming, *Lemmatizing*, das *Stopword Removal* und weitere. Ziel dieses Schrittes soll es sein, alle Texte soweit wie möglich auf Geschmacksbegriffe zu reduzieren. Dieser Schritt ist ebenfalls Bestandteil voriger Arbeiten und wurde für diese Arbeit im Detail verfeinert (Schole, 2017b). Dabei muss nach jedem Schritt eine Betrachtung seines Einflusses auf die weiteren Ergebnisse erfolgen. Eine Betrachtung der verschiedenen Vorverarbeitungsmethoden und ihrer Einflüsse auf die Texte findet in Kapitel 2.5 statt. Die technische Umsetzung und Verfeinerung dieses Schritts ist in den Kapiteln 3.3 bis 3.7 beschrieben.

Der Schritt der Datentransformation teilt sich im KDD-Prozess für diese Arbeit in mehrere Unterschritte auf. Zunächst muss ein Wissensmodell geschaffen werden, welches die Ähnlichkeiten von Geschmacksbegriffen abbildet. Dieses Wissensmodell lässt sich unter anderem mit dem Word2Vec-Algorithmus generieren (Mikolov u. a., 2013b). Durch Anwendung dieses Algorithmus auf die Tasting Notes des Trainingsdatensets entsteht ein Datenset aus Word Embeddings, welches die Ähnlichkeiten von Begriffen aus der Domäne Whisky abbildet. Die Eigentliche Datentransformation findet statt, indem die Word Embeddings auf die Tasting Notes des Testdatensets angewandt werden. Dies geschieht durch die Ermittlung der Word Embeddings der Begriffe, welche in einer Tasting Note vorkommen. Die Kombination der Word Embeddings einer Tasting Note durch die Berechnung des Mittelwerts bildet dann eine Repräsentationsform des Whiskys. Whiskys mit ähnlichen Geschmacksbegriffen in ihren Tasting Notes erhalten dabei nahe beieinanderliegende Vektoren.

Die Generierung der Word Embeddings mit dem Word2Vec-Algorithmus bietet einige Möglichkeiten zur Parametrisierung. Zudem hat die Vorverarbeitung der Texte einen großen Einfluss auf die resultierenden Embeddings. Daher ist es notwendig, an dieser Stelle verschiedene Möglichkeiten der Parametrisierung und Vorverarbeitung zu testen. Diese Wiederholung einzelner oder mehrerer Schritte ist Bestandteil des allgemeinen KDD-Prozesses und dient der Optimierung des Gesamtprozesses. Generell kann dieser Schritt auch als eigener KDD-Prozess zur Generierung eines Sets aus Word Embeddings, welches einen Geschmacksraum bildet, gesehen werden.

Auf dem entstandenen Set aus Whiskys und ihren zugehörigen repräsentativen Vektoren können bereits verschiedene Abfragen ausgeführt werden. Beispielsweise die nächsten Nachbarn eines Whiskys. Als Data-Mining-Methode findet in dieser Arbeit ein Clustering der Whisky-Vektoren statt. Die Repräsentationsform ist dabei durch die Whisky-Vektoren bereits gegeben. Als Distanzfunktion bieten sich unter anderem die euklidische und die Kosinus-Distanz an. Als Clustering-Methoden stehen viele Algorithmen zur Verfügung. Für diese Arbeit soll das

Clustering mit dem K-means-Algorithmus durchgeführt werden. Die so ermittelten Cluster bilden Geschmackskategorien im Testdatenset. Sowohl die Auswahl der nächsten Nachbarn als auch die ermittelten Cluster bilden eine Möglichkeit zur Validierung der Versuchsergebnisse. Eine mögliche Darstellungsform der Ergebnisse sind Visualisierungen der Embeddings und Cluster in Form von Grafiken. Eine Möglichkeit der Datenabstraktion ist es, zu jedem Cluster die am meisten auftretenden Begriffe zu ermitteln und diese als Repräsentationsform zu verwenden. Weitere Mögliche Abstraktionsformen sind eine Auswahl der am nächsten am beziehungsweise am entferntesten vom Cluster-Zentrum gelegenen Whiskys.

Im Rahmen der Evaluierung findet eine Expertenbefragung statt. Hierfür werden einerseits Sets von nächsten Nachbarn zu Referenz-Whiskys als Empfehlungen präsentiert und andererseits Auswahlen aus verschiedenen Clustern. Diese Whisky-Sets sollen die Experten dann im Bezug auf ihre Nachvollziehbarkeit bewerten.

3.2. Aufbau eines Datenkorpus

Als Grundlage für die weitere Arbeit muss zunächst ein möglichst umfangreicher Datenkorpus vorhanden sein. Dafür stehen mehrere verschiedene Arten von Ressourcen zur Verfügung. Nach einem Vergleich und der Entscheidung für eine Ressource müssen die Daten bezogen und in eine einheitliche Form gebracht werden. Im Gegensatz zu vorigen Arbeiten soll der Korpus in Form von einfachen TSV-Dateien gesichert werden. Neben den Tasting Notes sind auch die Metadaten zu den Whiskys von Interesse. Über diese lassen sich am Ende des KDD-Prozesses Vergleiche der Ermittelten Cluster mit realen Kategorien durchführen.

3.2.1. Recherche und Vergleich möglicher Datenquellen

Im Anschluss an die Recherche der Domäne Whisky mit ihren wichtigsten Merkmalen erfolgt eine Recherche möglicher Datenquellen für die Trainings- und Testdatensets. Aus vorigen Arbeiten sind bereits die Quellen Scotchwhisky.com (2018), Whisky.de (2018), Whisky Intelligence (2018), Whisky Magazine (2018) und Whisky Monitor (2018) bekannt, wobei Whisky Magazine (2018) als Testdatenset bestimmt wurde, da es sich hierbei um ein publizierendes Magazin handelt, dessen Tasting Notes von anerkannten Experten verfasst sind. Whisky.de (2018) hingegen wurde aus dem Datenset ausgeschlossen, da die Tasting Notes hier in der Regel nicht in Form von Texten, sondern in Form von quantitativen Werten dargestellt werden. Die übrigen Quellen dienen als Trainingsdatenset.

Im Rahmen der vorigen Arbeiten fiel die Entscheidung über die zu verwendenden Ressourcen auf Online-Ressourcen, da diese in einem deutlich höheren Maße verfügbar und leichter zu beziehen sind. Als einziges als Ressource geeignetes Buch wurde Murray (2016) ausgemacht.

Zu Beginn des neuen Durchlaufs des KDD-Prozesses erfolgt eine erneute Betrachtung der bisherigen Quellen und eine Recherche nach möglichen neuen Datenquellen. Die genannten Quellen bleiben in der Ressourcenliste erhalten und werden in diese Arbeit mit übernommen. Die Suche nach neuen Datenquellen erweist sich als schwierig, da potentiell geeignete Webseiten in der Regel von Whisky-Fans als Hobby betriebene Blogs sind und diese in gängigen Suchmaschinen über ein eher schlechtes Ranking verfügen. Einige der Hobby-Blogs bieten allerdings eine *Blogroll* an, über die weitere Blogs auffindbar sind. Im Zuge dieser Recherche wurden ergänzend zu den bisher verwendeten Quellen die Seiten Brossard (2018), Distiller (2018), Malt (2018), Master Of Malt (2018), Reddit (2018), Thomson (2018), Lardin (2018), Whisky Advocate (2018), Whisky Notes (2018), Whiskyology (2018) und Whiskey Reviewer (2018) entdeckt. Tabelle 3.1 zeigt eine Auflistung der verwendeten Ressourcen mit einer Einschätzung zur Professionalität und der Menge an verfügbaren Tasting Notes.

Ressource	Professionell / Hobby	Anzahl Tasting Notes
Distiller (2018)	professionell	2260
Scotchwhisky.com (2018)	professionell	926
Whisky Advocate (2018)	professionell	4298
Whisky Magazine (2018)	professionell	6952
Whiskyology (2018)	professionell	50
Brossard (2018)	Hobby	2155
Klaverstijn (2018)	Hobby	566
Malt (2018)	Hobby	162
Thomson (2018)	Hobby	1594
Lardin (2018)	Hobby	1199
Whisky Intelligence (2018)	Hobby	374
Whisky Monitor (2018)	Hobby	3120
Whisky Notes (2018)	Hobby	2853
Whiskey Reviewer (2018)	Hobby	891

Tabelle 3.1.: Datenquellen mit Einschätzung der Professionalität und Anzahl der verfügbaren Tasting Notes

Die Seiten Master Of Malt (2018) und Reddit (2018) wurden ebenfalls betrachtet, erwiesen sich jedoch als schwer zu beziehen. Daher wird auf diese Ressourcen aus zeitlichen Gründen

verzichtet. Die übrigen neuen Seiten wurden bezogen und in das Trainingsdatenset integriert. Die bereits bekannten Seiten wurden erneut bezogen, da gerade die professionell betriebenen Seiten regelmäßig neue Inhalte veröffentlichen. Generell muss erwähnt werden, dass Blogs, welche als Hobby betrieben werden, als Grundlage für Empfehlungen kritisch zu betrachten sind. Dies ist ein Grund, weshalb diese Quellen in jedem Fall nur für das Trainingsdatenset vorgesehen sind. Sie bilden eine Grundlage, um das Vokabular der Domäne zu lernen. Die professionell betriebenen Seiten bieten eine mögliche Grundlage für das Testdatenset. Für diese Arbeit soll dieses zunächst dennoch auf Whisky Magazine (2018) beschränkt werden.

3.2.2. Bezug und Pflege der ausgewählten Daten

Im Anschluss an die Recherche möglicher Datenquellen erfolgt die Entwicklung eines *Scrapers*, um Daten von diesen Ressourcen zu beziehen. Um die Daten automatisiert beziehen zu können, muss ein möglichst Variabler Scraper gebaut werden, welcher mit geringen Anpassungen auf neue Quellen angewandt werden kann. Diese Anpassungen sind im besten Fall ausschließlich über Konfigurationsdateien durchführbar.

Der Bezug der Daten fand im Rahmen vorangehender Arbeiten durch die Entwicklung angepasster Skripte zu jeder Ressource statt. Zudem wurden die Daten zunächst in einer CouchDB-Instanz und daraufhin in einer MySQL-Datenbank gespeichert (CouchDB, 2018; MySQL, 2018). Dies erwies sich aufgrund des Administrationsaufwandes allerdings als zu unpraktikabel, weshalb zu Beginn dieser Arbeit die Entscheidung steht, zur Speicherung der Daten einfache TSV-Dateien zu verwenden. Im Rahmen dieser Arbeit findet eine Neuentwicklung des Datengewinnungsprozesses statt. Dieser soll möglichst allgemein und variabel sein, um mit wenigen Anpassungen neue Quellen erschließen zu können. Eine Möglichkeit, Web Scraper variabel zu gestalten, ist die Verwendung von XPATH-Selektoren (Robie u. a., 2017). Diese ermöglichen es - ähnlich wie CSS-Selektoren (Glazman u. a., 2018) - , gezielt Elemente in einem HTML-Dokument auszuwählen. XPATH-Selektoren bieten dabei allerdings deutlich mehr Flexibilität. Ein derart gestalteter Web Scraper kann allein über die Anpassung von XPATH-Selektoren über eine Konfigurationsdatei auf das beziehen einer neuen Ressource eingerichtet werden.

Der Datengewinnungsprozess teilt sich in die Schritte Indizierung, Download und Parsen auf. Im Rahmen der Indizierung bezieht ein Skript sämtliche Links zu Detailseiten mit Tasting Notes der zu beziehenden Quelle. Bei der Entwicklung dieses Skripts fällt auf, dass die Quellen verschiedene Verzeichnisstrukturen verwenden. Professionelle Seiten bieten in der Regel eine Art Verzeichnis an. Dies kann auch in Form eines Menüs geschehen, wobei einige Seiten in

diesen auch weitere Kategorisierungen vornehmen. Dies macht die Fähigkeit des Skriptes, verschachtelte Verzeichnisse zu behandeln, notwendig. Blogs stellen ihre Inhalte in der Regel in Form einer Zeitleiste dar. Das heißt, die Artikel des Blogs werden chronologisch rückwärts angezeigt. Eine solche Zeitleiste kann wie ein mehrseitiges Verzeichnis behandelt werden. Wenige Seiten bieten eine einzelne Übersichtsseite mit allen Inhalten an. Das Indizierungsskript muss also verschiedene Strukturen behandeln können. Hierfür muss es entsprechend konfigurierbar entwickelt werden. So reicht es aus, in der Konfigurationsdatei festzulegen, auf welcher Seite mit der Indizierung begonnen werden soll, an welcher Stelle die gesuchten Link-Elemente zu finden sind, ob eine Kategorisierung vorliegt und somit auch Unterseiten betrachtet werden müssen und wo sich gegebenenfalls ein Button oder Link zur nächsten Seite des Verzeichnisses befindet. In Listing 3.1 ist die Vorgehensweise in Form eines Pseudocodes dargestellt.

```
1  browser.get(config.link)
2  categories = [config.link]
3  if config.is_nested:
4          categories = browser.find_all(config.categories_selector)
5
6  for category in categories:
7          stop = False
8          while not stop:
9                  browser.get(category)
10                 links = browser.find_all(config.links_selector)
11                 indexfile.write(links)
12                 if config.next_selector:
13                         browser.find(config.next_selector).click()
14                 else:
15                         stop = True
```

Listing 3.1: Pseudocode zur Indizierung von Webseiten

Die ermittelten Links zu Detailseiten speichert das Indizierungsskript in einer TSV-Datei ab. Diese verwendet das Download-Skript im folgenden Schritt, um sämtliche in der Datei gelisteten Links aufzurufen und die darüber erreichte Seite als HTML-Datei abzuspeichern. Die Speicherung geschieht lediglich aus Gründen der Performanz und Verfügbarkeit. Dadurch ist unter anderem sichergestellt, dass Internet-Ausfälle oder eine mögliche zwischenzeitliche Nichtverfügbarkeit einer Ressource die weitere Arbeit mit den Rohdaten verhindert.

Den letzten Schritt der Datengewinnung bildet das Parsen der gewonnenen HTML-Dokumente. Dies kann gleichzeitig bereits als erster Schritt der Datentransformation betrachtet werden,

da die Rohdaten hier in eine für Maschinen leichter zu verarbeitende Form gebracht werden. Ziel ist es, sämtliche verfügbaren Daten aus den Dokumenten, die aus einer Ressource geladen wurden, zu ermitteln und in einer strukturierten Form abzulegen. Vor dem Parsen ist es allerdings notwendig, die Indexdatei zu säubern. Diese kann je nach Ressource doppelte Einträge enthalten. Dies würde dazu führen, dass die betroffenen Whiskys doppelt erfasst würden. Zudem ist es möglich, dass einige ermittelte Links nicht erreichbar oder fehlerhaft sind. Diese müssen ebenfalls ausgeschlossen werden. Jede Ressource stellt unterschiedliche Metadaten zu den Whiskys bereit. Teilweise sind einige Metadaten nur durch die Anwendung von regulären Ausdrücken beispielsweise auf den Namen eines Whiskys zu ermitteln. In der Regel betrifft dies das Alter. Von zentralem Interesse ist immer die Tasting Note. Einige Ressourcen stellen mehrere Whiskys auf einer Seite dar. In diesen Fällen kommt es vor, dass keine strukturelle Unterscheidung zwischen den Whiskys vorliegt. Dies lässt sich dadurch erklären, dass die Inhalte des entsprechenden Blogposts ohne feste Struktur in einem einzigen *Richtext*-Feld gepflegt werden. Dies kann dazu führen, dass das Skript beim Parsen nicht in der Lage ist, sämtliche Sonderfälle abzudecken. In einem solchen Fall muss auf einen Teil der Texte der betroffenen Ressource verzichtet werden. Beim Parsen wird zwischen Metadaten, Tasting Notes und Ratings unterschieden. Zu einem Whisky können mehrere Tasting Notes und Ratings existieren. Listing 3.2 zeigt einen Pseudocode, welcher die Vorgehensweise beim Parsen verdeutlicht.

```
for filename in index_file:
        html = open(filename)
        for whisky in html.find_all(config.whisky_selector):
                metadata_file.write(read_metadata(whisky))
                for tasting_note in read_tasting_notes(whisky):
                        tasting_notes_file.write(tasting_note)
                for rating in read_ratings(whisky):
                        ratings_file.write(rating)
```

Listing 3.2: Pseudocode zum parsen von HTML-Dateien

Die Umsetzung der beschriebenen Skripte erfolgt in der Programmiersprache Python. Dabei findet vor allem die Bibliothek Selenium (2018) Verwendung, um einen *headless* Browser zu betreiben. Letzteres ist im Parse-Schritt nicht notwendig, da die zu parsenden Dateien lokal abliegen. In diesem Fall findet die Bibliothek *lxml* Verwendung, welche die Verwendung von XPath-Selektoren ermöglicht (Behnel u. a., 2018). Diese Bibliothek unterstützt lediglich XPath in Version 1. Dadurch ist die Funktionalität in einigen Details eingeschränkt. Am Ende dieses Prozesses steht ein Set aus Metadaten, Tasting Notes und Ratings in strukturierter Form.

Die einzelnen Datensätze können über IDs einander zugeordnet werden. Die verwendeten Datenquellen sind in Tabelle 3.1 in Kapitel 3.2.1 genannt.

Eine Besonderheit bildet die Ressource (Brossard, 2018). Diese scheint lediglich aus statischen HTML-Seiten zu bestehen und erweist sich dadurch als schwierig parsebar. Der Aufbau der Seiten unterscheidet sich mitunter stark. Aus diesem Grund werden aus dieser Ressource lediglich Tasting Notes ohne Metadaten bezogen. Die Ressource Whiskyology (2018) bietet eigentlich eine zu kleine Datenmenge. Das Parsen hätte keinen entsprechenden Mehrwert für das Datenset. Da die Daten der Seite allerdings über eine API im JSON-Format verfügbar sind, wurden sie über diesen Weg mit aufgenommen.

Zur Pflege gehört es, nach der Durchführung des Web-Scraping-Processes die gewonnenen Daten noch einmal zu bereinigen. Teilweise kann es vorkommen, dass Tasting Notes lediglich aus Links zu anderen Webseiten bestehen. In anderen Fällen sind die Tasting Notes leer.

3.3. Erster Durchlauf der Datenaufbereitung

Vor dem Data Mining müssen die Daten in eine geeignete Form gebracht werden. Diese Form soll ein Set aus Vektoren sein, welche die Verortung der Whiskys in einem geschmacklichen Raum repräsentieren. Hierfür muss zunächst ein Set aus Vektoren generiert werden, welche die einzelnen Aromen in einem geschmacklichen Raum verorten. Dies lässt sich durch die Anwendung des in Kapitel 2.6 beschriebenen Word2Vec-Algorithmus auf das Trainingsdatenset erreichen.

Während der verschiedenen Durchläufe der Datenaufbereitung werden einzelne, zufällig ausgewählte Texte aus dem Datenset betrachtet und dabei die Effekte der jeweils angewandten Vorverarbeitungsmethoden analysiert. Im Folgenden sind diese Beispieltexte zum Vergleich in ihrer originalen Form gezeigt.

> *„Nose: Notably sweet, but also tea-like.*
> *Palate: Satiny. Starts drier. More tightly combined flavours take longer to unfold. Gradually sweeter and more orangey (or tropical fruit?), but always with burnt-grass peatiness behind. Much more complex.*
> *Finish: Very long, soothing, warming.“*
>
> *Whyte & Mackay 18 Years Old* (Whisky Magazine, 2018)

> *„Nose: Soft, almost oily, mature mix of old column still rum, some char alongside white chocolate, ripe banana/banana chews and red fruit. Greener with water.*

> *Palate: Sweet start with those enthusiastic rummy elements, and a firm back-palate. Light chocolate (darker now) alongside raspberry. Soft, sweet and mature, but there's power here. Water reduces the complexity a little.*
> *Finish: Slightly sharp.“*
>
> *Invergordon 1997 Tiramisu Layers* (Scotchwhisky.com, 2018)

> *„The aroma is full of tropical golden fruits with a fairly decent intensity. Toasted coconut and sweet curry spices join in the fun. The palate shows much of the same with a nice balance between malt and fruit with the sherry casks heightening and not overwhelming the whisky. A trace of char smoke and salt is present, but only just. The finish is lovely and lasting.“*
>
> *Benromach 1976* (Distiller, 2018)

In Anhang A befinden sich neben diesen Texten auch Beispieltexte nach den einzelnen Verarbeitungsschritten aus allen Datenquellen.

3.3.1. Generierung von Word Embeddings

Als Grundlage für die Empfehlungen soll ein Set aus Word Embeddings dienen, in welchem Whisky-typische Geschmacksbegriffe nahe beieinander liegen. Dies soll über ein Training auf den erlangten Tasting Notes geschehen. Dabei gilt es allerdings, einige mögliche Parameter des Algorithmus zu beachten. Die Funktionalität des Algorithmus ist in Kapitel 2.6 beschrieben. Die Parameter Kontext- und Vektorgröße sind dort bereits erwähnt. Ein weiterer Parameter ist die Anzahl der zu durchlaufenden Trainingsepochen. Eine zu niedrige Anzahl führt zu ungenaueren Ergebnissen, während eine zu hohe Anzahl zu einer unnötig hohen Laufzeit ohne Verbesserung der Embeddings führt. Weiterhin hat die Größe des verwendeten Datensatzes einen entscheidenden Einfluss auf das Ergebnis. In dieser Arbeit findet die Implementierung des Word2Vec Algorithmus aus der Gensim-Bibliothek Anwendung (Řehůřek und Sojka, 2010). Zunächst soll dafür die Standard-Parametrisierung verwendet werden. Dies beinhaltet vorläufig auch die Verwendung des CBOW-Algorithmus im Kern.

Für die weitere Arbeit muss sichergestellt sein, dass die generierten Embeddings Qualitativ möglichst hochwertig sind. Ein Qualitätsindikator ist das Maß der Übereinstimmung mit bestehenden Strukturen. Als Vergleich kann hier ein Tasting Wheel dienen. Zudem kann nach Durchführung eines Clusterings ein Vergleich der ermittelten Cluster mit den realen Kategorien

aus den Tasting Wheels erfolgen. Allgemein stellt die Evaluierung der Embeddings eine große Herausforderung dar, da keine domänenspezifischen Vergleichsdatensets verfügbar sind.

Eine einfache, aber eher ungenaue Möglichkeit zur Bewertung der Word Embeddings ist die Visualisierung und Betrachtung dieser. Die Wortvektoren haben eine Größe von mehreren hundert Features. Daher müssen sie zunächst auf zwei Dimensionen reduziert werden, um sie optisch darstellen zu können. Eine Möglichkeit, ein Set aus multidimensionalen Vektoren auf zwei Dimensionen zu reduzieren, ist der *t-SNE*-Algorithmus (Maaten und Hinton, 2008). Dieser errechnet zu jedem Vektor eine zweidimensionale Repräsentation basierend auf den Entfernungen der Vektoren untereinander im originalen Raum. Dabei können unterschiedliche Distanzmetriken verwendet werden. Auf diese Weise lassen sich die Distanzverhältnisse der Embeddings untereinander Visualisieren. Zu beachten ist dabei, dass die Distanzen der resultierenden Abbildung nicht genau die Distanzen im ursprünglichen Raum wiedergeben, sondern lediglich Nachbarschaften darstellen. In dieser Arbeit findet die Implementierung dieses Algorithmus' aus der Scikit-Learn-Bibliothek Verwendung (Pedregosa u. a., 2011). Zur Erstellung der Grafiken dient die Bibliothek *Matplotlib* (Hunter, 2007).

Um den Word2Vec-Algorithmus verwenden zu können, muss diesem ein iterierbares Objekt übergeben werden. In jeder Iteration erwartet der Algorithmus einen Satz, welcher in Form einer Liste von Wörtern übergeben wird. Dies erfordert in jedem Fall ein erstes Preprocessing der Texte in dem diese zunächst in Sätze und dann weiter in Wörter aufgeteilt werden. Neben dieser Vorverarbeitung findet ebenfalls eine Konvertierung sämtlicher Buchstaben auf Kleinbuchstaben statt. Dies ist die naheliegendste erste Methode zur Angleichung verschiedener Texte. Zusätzlich zu den beiden genannten Methoden findet eine Entfernung von Satzzeichen statt. Dieses Verfahren ist für diese Arbeit geeignet, da grammatikalische Eigenschaften der Texte nicht von Relevanz sind.

Abbildung 3.1 zeigt die 200 häufigsten Wörter im Datenset ohne echtes vorhergehendes Preprocessing. Es wurden lediglich die genannten Verfahren angewandt. Die Embeddings sind mit der Standardparametrisierung des Word2Vec-Algorithmus generiert worden. Die Dimensionsreduktion erfolgt mit dem t-SNE-Alrogithmus mit den Parametern $perplexity = 50$ und $learning\_rate = 10$. Als Distanzmetrik dient die Kosinus-Distanz der Embeddings. Zur Initialisierung wird eine *Principal component analysis* durchgeführt. Dies bietet eine höhere Reproduzierbarkeit der Ergebnisse. Es ist bereits erkennbar, dass sich Wörter mit wenig Informationsgehalt am unteren rand der Grafik sammeln. Weiter ist erkennbar, dass einige übergeordnet beschreibende Begriffe zu Whiskys sich in der unteren linken Ecke anordnen. Geschmacksbe-

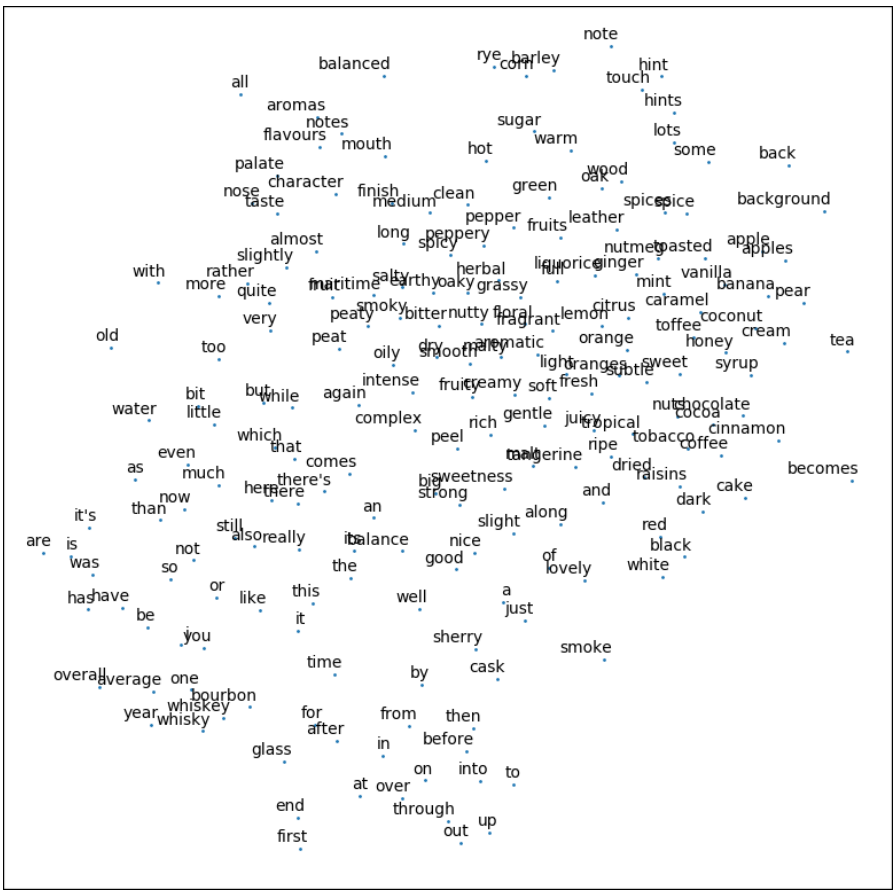


Abbildung 3.1.: Embeddings der 200 häufigsten Wörter im Datensatz ohne vorhergehendes Preprocessing

griffe tendieren nach rechts oben. Dabei lassen sich bereits Gruppierungen von unter anderem Farben, süßen Aromen und Gewürzen erkennen. Es ist allerdings auch erkennbar, dass die Entfernung von Stopwords vermutlich einen großen Einfluss auf die Ergebnisse hätte. Daher soll im Folgenden eine Erweiterung des Preprocessings um das Stopword-Removal stattfinden.

3.4. Zweiter Durchlauf der Datenaufbereitung

Nach der ersten Durchführung einer Transformation des Vokabulars des Textkorpus in Word Embeddings und der anschließenden Beurteilung der Ergebnisse findet ein Rückschritt zum

Preprocessing der Texte Statt. Dieses soll um ein Stopword-Removal erweitert werden. Dies wird zum Anlass genommen, das Preprocessing von Texten genauer zu betrachten.

3.4.1. Erweiterung des Preprocessings um ein Stopword-Removal

Die Vorverarbeitung von Texten hat zum Ziel, diese0 in eine Form zu bringen, in welcher sie möglichst optimal weiterverarbeitet werden können. Ein Teilziel dieser Arbeit ist es, eine Kombination aus Vorverarbeitungsmethoden zu ermitteln, die dies erfüllt.

Die Entfernung von Stopwords ist eine der am häufigsten verwendeten Methoden in der Vorverarbeitung von Texten. Für diese Arbeit dient sie als Annäherung an eine Reduktion der Texte auf Geschmacksbegriffe. Die Entfernung der Stopwords erfolgt unter Verwendung der Stopword-Liste der Python-Bibliothek *NLTK* (Bird und Loper, 2004). Listing 3.3 zeigt ein Minimalbeispiel für die Filterung von Stopwords aus einem Text.

```
>>> from nltk.corpus import stopwords
>>> sentence = "Toasted coconut and sweet curry spices "
... "join in the fun."
>>> sentence_without_stopwords = [
...     word for word in sentence.split()
...     if word not in set(stopwords.words('english'))
... ]
>>> str.join(" ", sentence_without_stopwords)
'toasted coconut sweet curry spices join fun.'
```

Listing 3.3: Beispielcode zur Entfernung von Stopwords

Die Verwendung der Stopwords in Form eines einfachen Sets ermöglicht es, dieses beliebig zu erweitern. Im Folgenden sind die Beispieltexte nach einer Entfernung der Stopwords angezeigt.

> *„notably sweet also tea like. satiny. starts drier. tightly combined flavours take longer unfold. gradually sweeter orangey tropical fruit. always burnt grass peatiness behind. much complex. long soothing warming.“*

> *„soft almost oily mature mix old column still rum char alongside white chocolate ripe banana banana chews red fruit. greener water sweet start enthusiastic rummy elements firm back palate. light chocolate darker alongside raspberry. soft sweet mature there's power. water reduces complexity little slightly sharp.“*

> *„aroma full tropical golden fruits fairly decent intensity. toasted coconut sweet curry spices join fun. palate shows much nice balance malt fruit sherry casks heightening overwhelming whisky. trace char smoke salt present. finish lovely lasting.“*

Die reduzierten Texte zeigen bereits eine Annäherung an die Wunschform. Dennoch verbleiben einige Wörter im Korpus, die nicht erwünscht sind. Eine Möglichkeit, diese Wörter ebenfalls zu entfernen ist es, die Stopword-Liste zu erweitern. Allerdings ist dies mit einem unverhältnismäßig großen Aufwand verbunden und führt nur mit sehr geringer Wahrscheinlichkeit zu einer Liste, die den Anspruch erfüllt, sämtliche unerwünschte Wörter zu beinhalten.

Entfernung von Wörtern unter Verwendung einer Whitelist

Eine Alternative zum Stopword-Removal ist die Verwendung einer Whitelist. Die Verwendung einer Whitelist, welche ausschließlich Geschmacksbegriffe enthält, ermöglicht die Reduzierung aller Texte auf ausschließlich erwünschte Wörter. Anhand der Beispieltexte bestünde eine solche Liste in etwa aus den Folgenden Wörtern:

> *sweet, tea, satiny, drier, tightly, sweeter, orangey, tropical, fruit, burnt, grass, peatiness, long, soothing, warming, soft, oily, mature, old, rum, char, chocolate, ripe, banana, chews, red, greener, water, rummy, firm, darker, raspberry, complexity, sharp, aroma, golden, fruits, intensity, toasted, coconut, curry, spices, malt, sherry, cask, smoke, salt, lovely, lasting*

Listing 3.4 zeigt eine Möglichkeit, Texte unter Verwendung einer Whitelist auf erwünschte Wörter zu reduzieren.

```
>>> import nltk
>>> import string
>>> sentence_tokens = nltk.sent_tokenize(text)
>>> sentences = []
>>> for sentence in sentence_tokens:
...     words = nltk.word_tokenize(sentence.replace('-', ' ')\
...     .translate(str.maketrans('', '', string.punctuation)))
...     sentences.append(words)
...
>>> whitelisted_only = [
...     [word for word in sentence if word in whitelist]
...     for sentence in sentences
... ]
>>> results = str.join(". ", [
...     str.join(" ", sentence) for sentence in whitelisted_only
... ]) + "."
```

Listing 3.4: Beispielcode zur Entfernung von Wörtern über eine Whitelist

Nach Anwendung der Whitelist auf die Beispieltexte sehen diese wie folgt aus:

> *„sweet tea. satiny. drier. tightly. sweeter orangey tropical fruit. burnt grass peatiness. long soothing warming.“*

> *„soft oily mature old rum char chocolate ripe chews red fruit. greener water. sweet rummy firm. chocolate darker raspberry. soft sweet mature. water complexity. sharp.“*

> *„aroma tropical golden fruits intensity. toasted coconut sweet curry spices. malt fruit sherry. char smoke salt. lovely lasting.“*

Die Ergebnisse zeigen, dass die Verwendung einer Whitelist sehr verlässlich zu einer Reduktion der Texte auf gewünschte Begriffe führt. Allerdings ist eine erkennbare Schwäche die Notwendigkeit, alle grammatikalischen Formen eines Wortes in die Liste aufzunehmen. Um dies zu verhindern, bietet es sich an, die Wörter zunächst auf ihre Stämme zurückzuführen. Diese Methode ist in Kapitel 3.5.1 beschrieben. Das Hauptproblem bei der Anwendung einer Whitelist ist jedoch die Verfügbarkeit einer geeigneten Liste. Hierzu existieren einige mögliche Quellen. *FooDB* und *FlavorDB* sind Datenbanken zu Geschmäckern und Aromen (Wishart, 2014; Garg u. a., 2017). Über diese lässt sich automatisiert eine große Whitelist für Geschmacksbegriffe

bilden. Auf Basis dieser beiden Datenbanken ist eine Liste von 1486 Begriffen generiert worden. Im Folgenden sind die Beispieltexte nach Anwendung dieser Whitelist dargestellt.

> *„sweet tea. tropical fruit. burnt grass.“*

> *„soft almost oily mix old rum white chocolate ripe banana banana red fruit. water sweet rummy. light chocolate raspberry. soft sweet. water slightly sharp.“*

> *„tropical fruits. toasted coconut sweet curry spices. malt fruit sherry whisky. smoke salt.“*

Es ist erkennbar, dass das Verfahren beim zweiten und dritten Text einigermaßen gut funktioniert. Am ersten Text ist jedoch erkennbar, dass die Kürzung deutlich zu hoch ausgefallen ist. Speziell die Begriffe *satiny* und *drier*, welche das Mundgefühl beschreiben, fehlen. Aber auch die deutlich wichtigeren, Whisky-spezifische Geschmacksbegriffe wie *peatiness* sind entfernt worden. Damit lässt sich sagen, dass diese Methode für allgemeine Geschmacksbegriffe gut funktioniert. Bei einem vorher angewandten Stemming ist ebenfalls von besseren Ergebnissen auszugehen. Allerdings sind gerade Begriffe wie *peat* von großer Bedeutung in der Domäne Whisky. Zudem kann nicht gewährleistet werden, dass die Whitelist sämtliche Begriffe beinhaltet, die weiter verwendet werden sollen. Die manuelle Erweiterung dieser Liste birgt ein zu großes Risiko, nicht zu einer ausreichend vollständigen Liste zu führen. Eine Erweiterung der Whitelist um Begriffe aus Nosing Wheels garantiert ebenfalls nicht, dass diese anschließend vollständig ist. Daher wird im weiteren Verlauf dieser Arbeit zunächst das Stopword-Removal verwendet.

3.4.2. Generierung der Word Embeddings

Nach der Analyse der möglichen Verfahren zur Reduzierung der Texte auf Geschmacksbegriffe und der Festlegung auf das Stopword-Removal folgt ein erneutes Training der Word Embeddings. Abbildung 3.2 zeigt die häufigsten 200 Wörter im Datenset nach vorhergegangenem Stopword-Removal.

Abgesehen vom Fehlen der Stopwords ist erkennbar, dass die Gruppierungen enger und damit nachvollziehbarer sind. Die übergeordneten Begriffe sind klarer abgegrenzt. Rohstoffe wie *malt*, *barley*, *grain*, *rye* und *corn* liegen nahe beieinander. Weiterhin lassen sich feinere Gruppierungen erkennen. So sind die Begriffe für fruchtige Aromen unter anderem in Zitrusfrüchte und Obst unterteilt. Andere logisch zusammenhängende Begriffe wie *tongue*, *palate* und *mouth* liegen

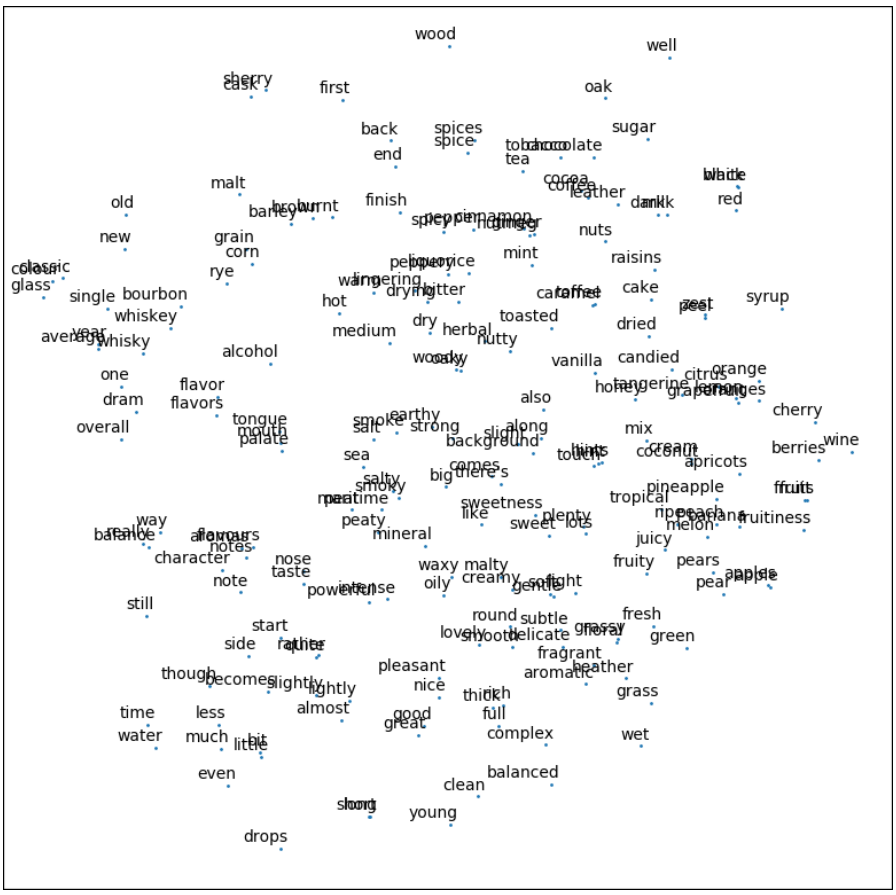


Abbildung 3.2.: Embeddings der 200 häufigsten Wörter im Datensatz nach Stopword-Removal

ebenfalls beieinander. Dennoch lässt sich erkennen, dass sich beispielsweise Adjektive, welche auf *-y* enden im Zentrum der Grafik sammeln. Zudem ergibt eine Unterscheidung zwischen Singular und Plural an dieser Stelle keinen Sinn. Beispiele hierfür sind *spice* → *spices* und *pear* → *pears* Diese grammatikalischen Einflüsse können über das Stemming eingedämmt werden. Daher soll im Folgenden der Einfluss dieses Verfahrens auf die Word Embeddings überprüft werden.

3.5. Dritter Durchlauf der Datenaufbereitung

Nach der Betrachtung des Einflusses des Stopword-Removals auf die Texte und Word Embeddings soll das Preprocessing noch weiter um das Stemming erweitert werden. Dies beschreibt die Reduzierung aller Wörter auf ihren Wortstamm.

3.5.1. Erweiterung des Preprocessings um ein Stemming

Das Preprocessing soll nach dem Entfernen unerwünschter Wörter um ein Stemming erweitert werden, da viele Wörter weiterhin in unterschiedlichen Schreibweisen und grammatikalischen Formen vorliegen. Es bestehen mehrere Möglichkeiten, unterschiedliche Formen desselben Begriffes zu vereinheitlichen. Die beiden am meisten verwendeten Verfahren sind das Stemming und das Lemmatizing. Stemming beschreibt die Rückführung von Wörtern auf ihren Stamm. Dabei muss das Ergebnis nicht der reale Wortstamm sein. Das Lemmatizing ermittelt zu jedem Wort unter Verwendung einer *Lookup-Table* eine Grundform. Es existieren einige verschiedene Stemming-Algorithmen. Bekannte Beispiele hierfür sind der Porter Stemmer, der Lancaster Stemmer und der Snowball Stemmer. Im folgenden sind die Beispieltexte nach Anwendung der drei Stemmer und dem WordNet Lemmatizer gezeigt (Miller, 1995). Die Wortreduktion findet dabei unter Anwendung der Implementierungen aus der Python Bibliothek NLTK statt (Bird und Loper, 2004).

Porter Stemmer

> *„notabl sweet also tea like. satini. start drier. tightli combin flavour take longer unfold. gradual sweeter orangey tropic fruit. alway burnt grass peati behind. much complex. long sooth warm.“*

> *„soft almost oili matur mix old column still rum char alongsid white chocol ripe banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there' power. water reduc complex littl slightli sharp.“*

> *„aroma full tropic golden fruit fairli decent intens. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“*

Lancaster Stemmer

„not sweet also tea lik. satiny. start dri. tight combin flavo tak long unfold. grad sweet orangey trop fruit. alway burnt grass peaty behind. much complex. long sooth warm.“

„soft almost oi mat mix old column stil rum char alongsid whit chocol rip banan banan chew red fruit. green wat sweet start enthusiast rummy el firm back pal. light chocol dark alongsid raspberry. soft sweet mat there's pow. wat reduc complex littl slight sharp.“

„arom ful trop gold fruit fair dec intens. toast coconut sweet curry spic join fun. pal show much nic bal malt fruit sherry cask height overwhelm whisky. trac char smok salt pres. fin lov last.“

Snowball Stemmer

„notabl sweet also tea like. satini. start drier. tight combin flavour take longer unfold. gradual sweeter orangey tropic fruit. alway burnt grass peati behind. much complex. long sooth warm.“

„soft almost oili matur mix old column still rum char alongsid white chocol ripe banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“

„aroma full tropic golden fruit fair decent intens. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“

WordNet Lemmatizer

„notably sweet also tea like. satiny. start drier. tightly combined flavour take longer unfold. gradually sweeter orangey tropical fruit. always burnt grass peatiness behind. much complex. long soothing warming.“

„soft almost oily mature mix old column still rum char alongside white chocolate ripe banana banana chew red fruit. greener water sweet start enthusiastic rummy element

firm back palate. light chocolate darker alongside raspberry. soft sweet mature there's power. water reduces complexity little slightly sharp.“

„aroma full tropical golden fruit fairly decent intensity. toasted coconut sweet curry spice join fun. palate show much nice balance malt fruit sherry cask heightening overwhelming whisky. trace char smoke salt present. finish lovely lasting.“

Ein Vergleich der Ergebnisse zeigt Schwächen und Stärken der einzelnen Algorithmen auf. Die Texte nach Anwendung des Lancaster Stemmers zeigen deutlich, dass dieser einen zu großen Teil der Worte entfernt. Teilweise gehen dadurch wichtige Informationen verloren, die die Wörter von anderen unterscheiden. Beispiele hierfür sind *notably* → *not* und *mature* → *mat*. In anderen Fällen wie *darker* → *dark* funktioniert der Lancaster Stemmer allerdings besser als die Vergleichs-Stemmer. Die Ergebnisse des Lemmatizers zeigen, dass dieser deutlich zu wenig von den Wörtern entfernt. Er bietet allerdings den Vorteil, dass die Wörter nicht in eine abstrakte Form gebracht werden. Tabelle 3.2 vergleicht die Ergebnisse der verschiedenen Stemmer anhand einiger Grenzfälle, die im Rahmen dieser Arbeit auftreten.

Die Tabelle zeigt deutlich, dass der WordNet Lemmatizer nicht in der Lage ist, Wörter aus diesem Beispielset in eine gleiche Form zu bringen. Ebenso ist erkennbar, dass alle drei Algorithmen Endungen wie *-ey* nicht behandeln. Eine Entfernung dieser Endung hätte zur Folge, dass Wörter wie *honey* ebenfalls beschnitten würden. Da Schreibweisen wie *smokey* oder *orangey* eher die Ausnahme bilden, ist dieser Nachteil zu vernachlässigen. Ein Vergleich der Stemming-Algorithmen zeigt, dass alle drei ähnlich gute Ergebnisse liefern. Porter und Snowball Stemmer stimmen dabei fast überein. Der Lancaster Stemmer entfernt wie bereits erwähnt teilweise zu viel Information. Die anderen beiden Stemmer führen hingegen die Wörter *nutty* und *nuttyness* nicht auf den selben Stamm zurück. Dies kann allerdings durch ein erneutes Anwenden des jeweiligen Stemmers auf das Wort erreicht werden. Der Snowball Stemmer behandelt die Endung *-ly* besser als der Porter Stemmer. Als am besten geeignete Methode erscheint hier also eine doppelte Anwendung des Snowball Stemmers.

Beispieltexte nach doppelter Anwendung des Snowball Stemmers

„notabl sweet also tea like. satini. start drier. tight combin flavour take longer unfold. gradual sweeter orangey tropic fruit. alway burnt grass peati behind. much complex. long sooth warm.“

Wort	Porter Stemmer	Lancaster Stemmer	Snowball Stemmer	WordNet Lemmatizer
nut	nut	nut	nut	nut
nutty	nutti	nutty	nutti	nutty
nuttyness	nutty	nutty	nutty	nuttyness
peat	peat	peat	peat	peat
peaty	peati	peaty	peati	peaty
peatiness	peati	peaty	peati	peatiness
spices	spice	spic	spice	spice
spicy	spici	spicy	spici	spicy
smoke	smoke	smok	smoke	smoke
smoky	smoki	smoky	smoki	smoky
smokey	smokey	smokey	smokey	smokey
fruits	fruit	fruit	fruit	fruit
fruity	fruiti	fru	fruiti	fruity
darker	darker	dark	darker	darker
dark	dark	dark	dark	dark
mature	matur	mat	matur	mature
maturity	matur	mat	matur	maturity
honey	honey	honey	honey	honey
slightly	slightli	slight	slight	slightly
tightly	tightli	tight	tight	tightly

Tabelle 3.2.: Vergleich von Beispielwörtern nach Anwendung von Porter Stemmer, Lancaster Stemmer und WordNet Lemmatizer

> *„soft almost oili matur mix old column still rum char alongsid white chocol ripe banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“*

> *„aroma full tropic golden fruit fair decent inten. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“*

An Tabelle 3.2 ist weiterhin Optimierungspotential zu erkennen. Begriffe wie *peat* und *peati* müssten in die gleiche Form gebracht werden. Dies könnte über ein zusätzliches Entfernen des Buchstabens i von Wortenden erreicht werden. Allerdings ist nicht vollständig ersichtlich,

ob dies in anderen Fällen zu unerwünschtem Informationsverlust führt. Gleiches gilt für die Begriffe *darker* und *dark*. Eine Reduktion des Wortes *nutty* auf *nut* ist ebenso wünschenswert. Aufgrund der erwähnten Unvorhersehbarkeit dieser Verfahren werden diese zunächst nicht weiter verfolgt.

3.5.2. Generierung der Word Embeddings

Nach der Betrachtung der verschiedenen Methoden zum Stemming und der Festlegung auf eine doppelte Anwendung des Snowball-Stemmers müssen die Word Embeddings erneut trainiert werden. Abbildung 3.3 zeigt die häufigsten 200 Wörter im Datenset nach vorigem Stopword-Removal und doppeltem Stemming.

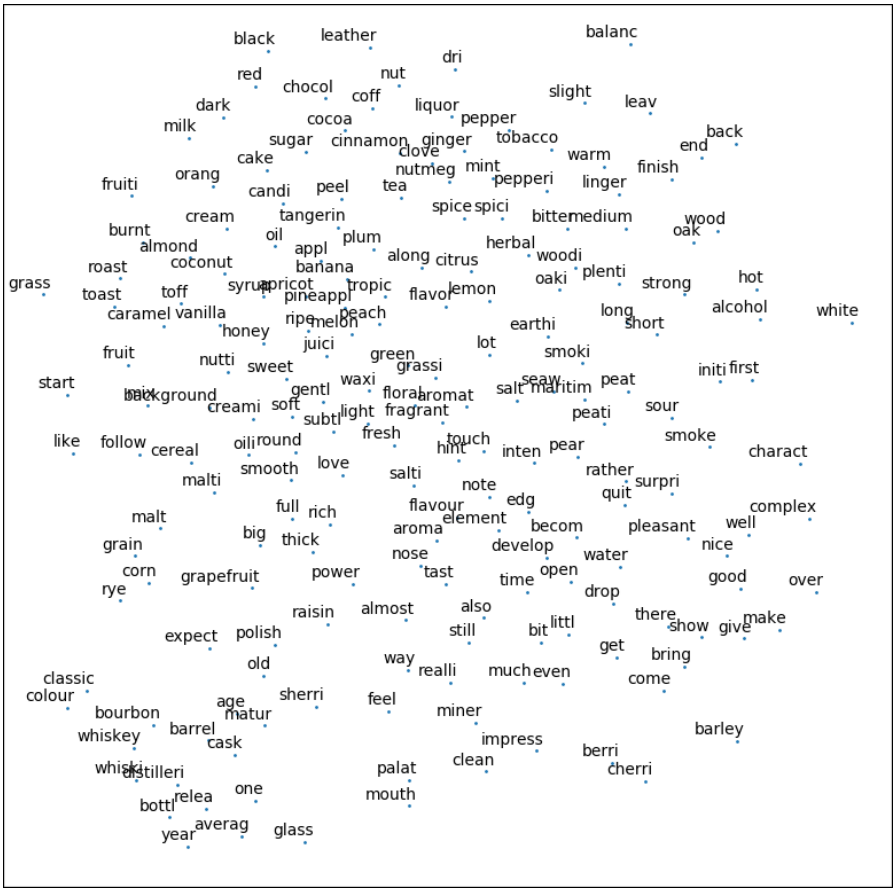


Abbildung 3.3.: Embeddings der 200 häufigsten Wörter im Datensatz nach Stemming

Ein großer Unterschied in der Aufteilung der Wörter ist nicht direkt erkennbar. Generell sind die Wörter deutlich gleichmäßiger verteilt. Dies hat allerdings vermutlich seinen Ursprung in der Beschaffenheit des *t-SNE*-Algorithmus. Es finden sich keine Plurale mehr in der Darstellung, was ein Ziel des Stemmings ist. Die Adjektive lassen sich über das Stemming allerdings nicht entfernen. An dieser Grafik ist weiterhin zu erkennen, dass einige Farben in den häufigsten Wörtern vorkommen. Diese haben alleinstehend weniger Informationsgehalt als in Kombination mit dem jeweils von ihnen Beschriebenen Begriff. Daher soll im Folgenden eine Phrasenerkennung durchgeführt werden.

3.6. Vierter Durchlauf der Datenaufbereitung

Nach den Erweiterungen des Preprocessings um ein Stopword-Removal und ein Stemming besteht weiterhin das Problem, dass Begriffe, welche sich aus mehreren Wörtern zusammensetzen, nicht als solche behandelt werden. Dies führt dazu, dass beispielsweise alleinstehende Farben Informationen aus verschiedenen Kontexten in sich tragen.

3.6.1. Erweiterung des Preprocessings um eine Phrase Detection

Viele Begriffe bestehen aus mehr als einem Wort. Die Bedeutung der beiden Wörter in Kombination kann dabei stark von der Bedeutung der einzelnen Wörter abweichen. Diese Wortkombinationen müssen also vor der weiteren Verarbeitung erkannt und als zusammenhängende Begriffe markiert werden. Eine Methode, um zusammengehörende Begriffe zu ermitteln ist die Phrase Detection. Diese ermittelt zu jedem Wortpaar die Häufigkeit des gemeinsamen Auftretens im Verhältnis zum Korpus. Dabei können verschiedene Metriken angewandt werden. Für die Phrase Detection findet hier die Python-Bibliothek Gensim Anwendung (Řehůřek und Sojka, 2010). Listing 3.5 zeigt einen Minimalcode zur Phrasenerkennung in einem Korpus.

```
>>> from gensim.models.phrases import Phrases, Phraser
>>> phrases = Phrases(sentences)
>>> phraser = Phraser(phrases)
>>> sentence = ['orang', 'peel', 'should', 'be', 'recognized', 'as',
... 'phrase', 'black', 'pepper', 'too']
>>> phraser[sentence]
['orang_peel', 'should', 'be', 'recognized', 'as', 'phrase',
```

```
'black_pepper', 'too']
```

Listing 3.5: Beispielcode zur Erkennung von Phrasen in einem Textkorpus

Beispieltexte nach Anwendung der Phrase Detection

> *„notabl sweet also tea like satini. start drier. tight_combin flavour take longer unfold. gradual sweeter orangey tropic_fruit. alway burnt_grass peati behind. much complex. long sooth warm.“*

> *„soft almost oili matur mix old column_still rum char alongsid white_chocol ripe_banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“*

> *„aroma full tropic golden fruit fair decent inten. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri_cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“*

Die Betrachtung der Texte zeigt einige sinnvoll erkannte Begriffe und einige weniger sinnvolle. Die Begriffe *tropic fruit*, *burnt grass*, *white chocol*, *ripe banana* und *sherri cask* sind sinnvolle Begriffe, die Kombination *tight combin* weniger. Die Wortkombinationen *red fruit* und *back palat* sind Beispiele für möglicherweise ebenfalls zu erkennende Begriffe.

Eine Möglichkeit, den Phrasenerkennungsalgorithmus anzupassen, ist die Verwendung einer anderen Scoring-Funktion. Diese Funktion errechnet zu jedem Wortpaar einen Wert, der seine Eignung als Phrase beschreibt. Überschreitet der Wert einen Schwellwert, wird das entsprechende Wortpaar als Phrase erkannt. Dieser Schwellwert ist ebenfalls anpassbar.

$$score(a,b) = \frac{(häufigkeit\_wortpaar - min\_count) \cdot grösse\_vokabular}{häufigkeit\_a \cdot häufigkeit\_b} \tag{3.1}$$

$$score(a,b) = \frac{\ln(P(a,b)/(P(a) \cdot P(b)))}{-\ln(P(a,b))},\ wobei\ P(wort) = \frac{häufigkeit\_wort}{grösse\_vokabular} \tag{3.2}$$

Gleichung 3.1 zeigt die Standardfunktion zur Ermittlung eines Scores für ein Wortpaar *a* und *b*. Dabei ist *min_count* ebenfalls ein variabler Parameter. Wortpaare, die seltener als *min_count* vorkommen, werden nicht berücksichtigt. Gleichung 3.2 zeigt eine alternative Funktion.

Beispieltexte nach Anwendung der Phrase Detection mit alternativer Scoring-Funktion

> *„notabl sweet also tea like satini. start drier. tight combin flavour take longer unfold. gradual sweeter orangey tropic_fruit. alway burnt grass peati behind. much complex. long sooth warm.“*

> *„soft almost oili matur mix old column still rum char alongsid white chocol ripe banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“*

> *„aroma full tropic golden fruit fair decent inten. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“*

Die Beispiele zeigen, dass deutlich weniger Begriffe ermittelt werden. Dies kann an einem zu hohen Wert für den Schwellwert liegen. Dieser liegt für die obigen Beispiele bei 0.5. Im Folgenden sind die Texte nach einer Verringerung des Wertes auf 0.4 gezeigt.

> *„notabl sweet also tea like satini. start drier. tight_combin flavour take longer unfold. gradual sweeter orangey tropic_fruit. alway burnt grass peati behind. much complex. long sooth warm.“*

> *„soft almost oili matur mix old column_still rum char alongsid white_chocol ripe_banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“*

> *„aroma full tropic golden fruit fair decent inten. toast coconut sweet curri spice join_fun. palat show much nice balanc malt fruit sherri_cask heighten_overwhelm whiski. trace char smoke salt present. finish love last.“*

Diese Texte sind bereits deutlich näher an den ermittelten Phrasen unter Verwendung der Standardfunktion. Es lässt sich dennoch keine Qualitätssteigerung beobachten, welche einen Wechsel der Funktion rechtfertigen würde. Da die Standardfunktion schneller zu berechnen ist, findet diese im weiteren Verlauf Verwendung.

Mehrfache Anwendung der Phrase Detection

Viele Begriffe setzen sich aus mehr als zwei Wörtern zusammen. Um solche Wortkombinationen ebenfalls zu erfassen, besteht die Möglichkeit, das Phrasing mehrmals durchzuführen. Durch eine doppelte Anwendung der Phrasenerkennung sind bereits Phrasen, welche aus vier Wörtern bestehen, ermittelbar. Im Folgenden sind die Beispieltexte nach doppelter Phrasenerkennung dargestellt.

> *„notabl sweet also tea like satini. start drier. tight_combin_flavour take longer unfold. gradual sweeter orangey tropic_fruit. alway burnt_grass peati behind. much complex. long sooth warm.“*

> *„soft almost oili matur mix old column_still rum char alongsid white_chocol ripe_banana banana chew red_fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.“*

> *„aroma full tropic golden fruit fair decent inten. toast_coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri_cask heighten overwhelm whiski. trace char smoke salt present. finish love last.“*

An den Beispielen ist lediglich erkennbar, dass die Wortkombination *tight combin flavour* als Begriff erfasst wurde. Dies lässt sich darauf zurückführen, dass die einzelnen Worte dieser Gruppe im restlichen Textkorpus nicht häufig vorkommen. In diesem Fall bietet die Phrase Detection eine Möglichkeit, seltene Wörter bereits in diesem Schritt auszuschließen. Zudem ermöglicht die Phrase Detection eine Erkennung häufig auftretender Folgen von Wörtern. Diese können beliebte Phrasen eines bestimmten Autors sein. Solche Phrasen könnten möglicherweise Entfernt werden, um den Einfluss einzelner Autoren zu verringern. Weiter ist an den Texten erkennbar, dass die Kombindation *red fruit* bei der zweiten Anwendung der Phrase Detection als Begriff erkannt wird. Es liegt nahe, dass das Wort *fruit* durch die erste Anwendung bereits mit anderen Wörtern kombiniert wurde. Dadurch hat sich die Häufigkeit des Wortes alleine verringert, was den Score für die Kombination *red fruit* erhöht. Gleiches ist für das Wort *red* denkbar. Insgesamt kann hier von einer Verbesserung gesprochen werden, weshalb die doppelte Anwendung der Phrase Detection beibehalten wird.

3.6.2. Generierung der Word Embeddings

Nach der Analyse mehrerer Möglichkeiten zur Paremetrisierung der Phrase Detection muss diese nun auf das Trainingsset angewandt und anschließend die Ergebnisse erneut betrachtet werden. Abbildung 3.4 zeigt die häufigsten 200 Wörter nach Stopword-Removal, Stemming und anschließender doppelter Phrase Detection. Es lässt sich erkennen, dass einige Begriffe wie *tropic fruit* und *dark chocolate* korrekt ermittelt werden und diese auch passend Positioniert sind. Weiter lässt sich eine deutlichere Aufteilung der Wörter in Aromen und andere beschreibende Begriffe beobachten. Die Geschmacksbegriffe bilden eine große, ovale Anordnung während die beschreibenden Begriffe eher außerhalb dieses Bereichs angesiedelt sind.

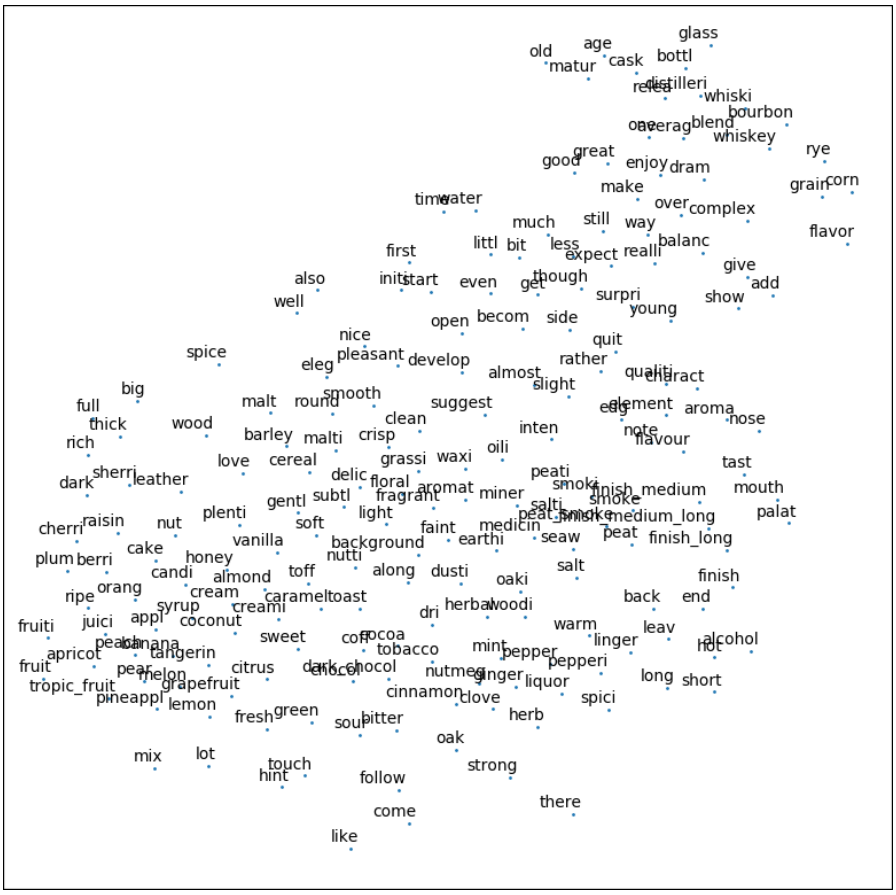


Abbildung 3.4.: Embeddings der 200 häufigsten Wörter im Datensatz nach Phrase Detection

Nach der Betrachtung der Phrase Detection und ihres Einflusses auf die Word Embeddings findet zunächst kein weiteres Preprocessing der Texte statt. Stattdessen soll im folgenden die Parametrisierung des Trainingsprozesses optimiert werden.

3.6.3. Optimierung des Wortvektortrainings

Das Training der Word Embeddings bietet einige Möglichkeiten zur Parametrisierung, welche an dieser Stelle Untersucht werden sollen. Hierfür soll zunächst eine Betrachtung der aktuell vorhandenen Word Embeddings stattfinden. Dies kann unter anderem über eine Visualisierung des gesamten Sets von Word Embeddings geschehen.

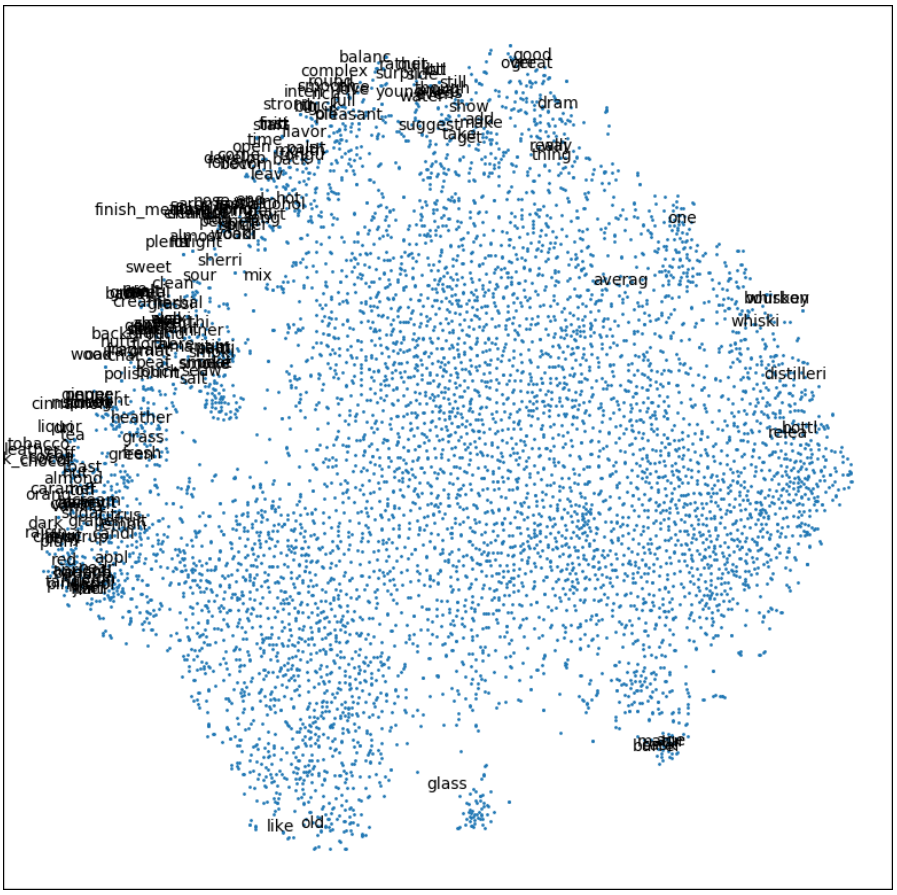


Abbildung 3.5.: Embeddings des gesamten Vokabulars nach Phrase Detection

Abbildung 3.5 zeigt im Vergleich zu Abbildung 3.4 eine Darstellung des gesamten Vokabulars. Es ist erkennbar, dass die Häufigsten Wörter sich im Vergleich zum restlichen Datenset am Rand befinden. Der Abstand zwischen Geschmacksbegriffen und übergeordneten Begriffen ist in dieser Darstellung deutlich größer als in den vorigen.

Als unterstützendes Mittel zur Bewertung der Embeddings eignet sich ein Clustering dieser. Die dabei ermittelten Cluster bieten einen Eindruck der Verteilung der Begriffe im multidimensionalen Raum. Für das Clustering der Embeddings bieten sich verschiedene Algorithmen an. Zunächst soll hierfür der K-means-Algorithmus angewandt werden. Dieser gilt als wichtigster flacher Clustering-Algorithmus (Schütze u. a., 2008, S. 360). In diesem Fall findet die Implementierung aus der Scikit-Learn-Bibliothek Verwendung (Pedregosa u. a., 2011). Der K-means-Algorithmus erfordert eine vorgegebene Anzahl zu ermittelnder Cluster. Zu Beginn des Trainings bestimmt K-means zufällige Cluster-Zentren. Diesen ordnet der Algorithmus im folgenden Schritt sämtliche Elemente im Datenset anhand ihrer Distanz zu. Daraufhin erfolgt eine Neuberechnung der Cluster-Zentren basierend auf den Elementen der einzelnen Cluster. Die letzten beiden Schritte wiederholt der Algorithmus, bis die Summe der quadrierten Distanzen der einzelnen Datensätze zu ihren jeweiligen Cluster-Zentren ein Minimum erreicht. Dieser Wert trägt die Bezeichnung *Residual sum of squares*, im Folgenden RSS. K-means ist dabei stark von der Auswahl der initialen Cluster-Zentren abhängig. Um diesem entgegenzuwirken, wird der Algorithmus in der Regel mehrmals ausgeführt und das Ergebnis mit der niedrigsten RSS ausgewählt. Die hier verwendete Implementierung führt den Algorithmus zehn mal aus und stoppt jede Ausführung nach maximal dreihundert Iterationen. Eine Methode, die initiale Auswahl an Cluster-Zentren zu verbessern, ist die Weiterentwicklung *K-means++* (Arthur und Vassilvitskii, 2007). Diese wählt die Cluster-Zentren so aus, dass sie möglichst weit voneinander entfernt sind. Die Scikit-Learn-Implementierung des Algorithmus verwendet diese Methode standardmäßig. Da der K-means-Algorithmus allgemein schnell ist und die Geschwindigkeit des Clusterings für diese Arbeit generell nicht von Bedeutung ist, wird die Anzahl der Ausführungen des Algorithmus auf hundert und die maximale Anzahl an Iterationen pro Ausführung auf zehntausend erhöht. Diese Anpassung erfolgt zur Sicherstellung der bestmöglichen Ergebnisse.

Abbildung 3.6 zeigt das gesamte Vokabular nach einem Clustering mit der Standardanzahl an zu ermittelnden Clustern. In der Standardparametrisierung der verwendeten Implementierung ermittelt der Algorithmus acht Cluster. Es ist erkennbar, dass die am Rand liegenden Embeddings in gemeinsamen Clustern liegen und im Zentrum ein größerer Cluster liegt. Die Vermutung liegt nahe, dass diese Aufteilung aufgrund der Häufigkeit dieser Wörter geschieht. Der Word2Vec-

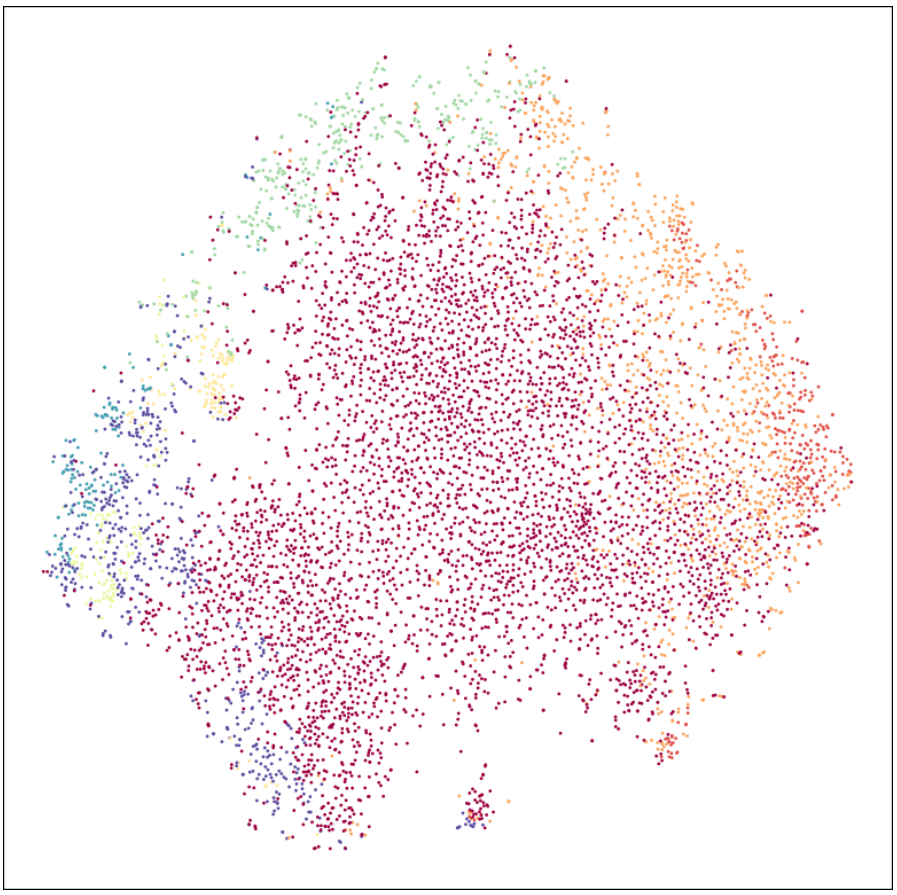


Abbildung 3.6.: Embeddings des gesamten Vokabulars, eingefärbt nach Clustern

Algorithmus hat die Eigenschaft, zu in mittlerer Häufigkeit auftretenden Wörtern längere Vektoren zu ermitteln. Besonders selten oder häufig auftretende Wörter erhalten dagegen kürzere Vektoren (Schakel und Wilson, 2015). Die Kürze der Vektoren besonders häufiger Wörter ist damit zu erklären, dass diese in vielen verschiedenen Kontexten vorkommen. Daher haben diese einen geringeren Informationsgehalt. Das Clustering ist in diesem Fall somit durch die Häufigkeit der Wörter beeinflusst. Anders formuliert stellen die Cluster eher eine Darstellung der Relevanz der Wörter dar. Gegebenenfalls bietet dieses Clustering auch eine Möglichkeit, zu seltene und zu häufige Wörter auszuschließen. Dies ist allerdings bereits durch andere Maßnahmen in vorigen Schritten der Datenverarbeitung möglich und wird daher an dieser Stelle zunächst nicht weiter verfolgt. Die Überschneidung der Cluster in der Grafik lässt sich

ebenfalls mit dem Einfluss der Häufigkeit einzelner Wörter erklären. Zudem findet hier eine Reduktion von Embeddings der Dimension hundert auf zwei Dimensionen statt. Dies führt zwangsläufig zu Ungenauigkeiten in der Darstellung. Der Einfluss der Vektorlänge und damit der Wortrelevanz lässt sich durch eine Normalisierung der Vektoren vor dem Clustering eliminieren.

Optimierung des Clusterings der Word Embeddings

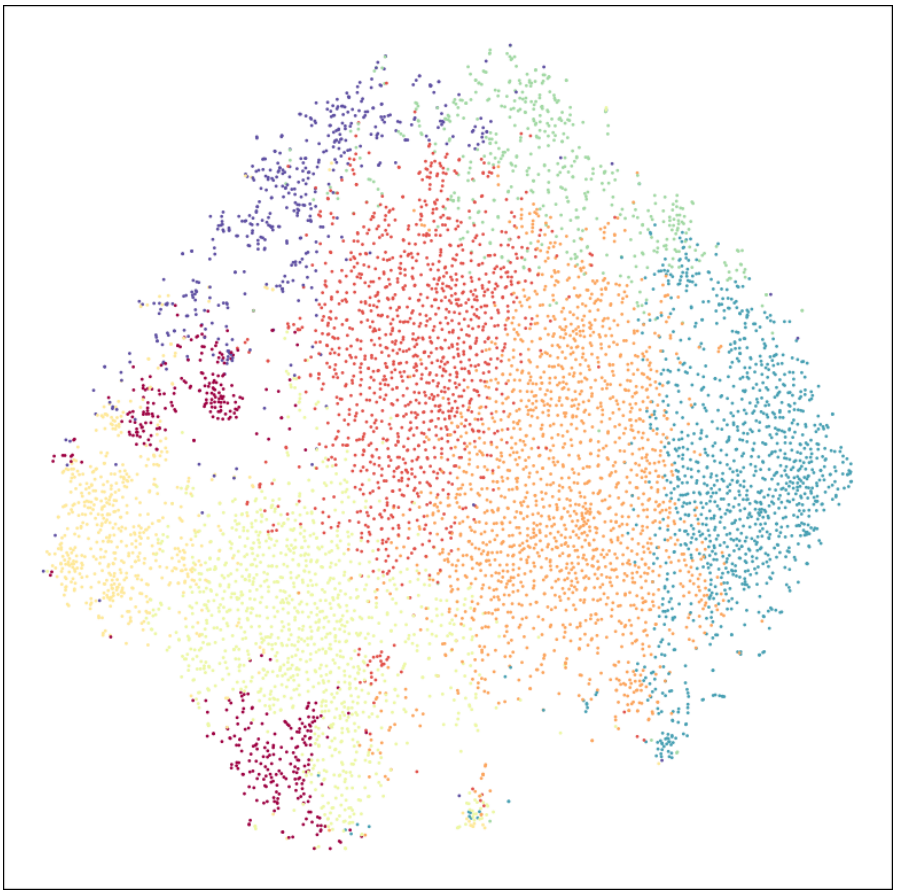


Abbildung 3.7.: Embeddings des gesamten Vokabulars, eingefärbt nach Clustering der normalisierten Embeddings

Abbildung 3.7 zeigt die Daten nach Clustering mit vorangegangener Normalisierung der Wortvektoren. Es ist auf den ersten Blick erkennbar, dass sich die Visualisierung und das Clustering erkennbar deutlicher miteinander decken. Zudem ist das Größenverhältnis der Cluster untereinander ausgeglichener.

Das Clustering bietet weiterhin Optimierungspotential. Eine Verbesserung der Ergebnisse kann besonders durch die Bestimmung der am besten geeigneten Anzahl an Clustern erreicht werden. Zum Vergleich mehrerer Clustering-Ergebnisse miteinander bieten sich verschiedene Faktoren an. Einer ist der Silhouettenkoeffizient. Dieser errechnet zu jedem Element o das Verhältnis seiner Distanz zum eigenen Cluster A zu der zum nächstgelegenen Cluster B. Gleichung 3.3 zeigt die Formel zur Berechnung der Silhouette eines Elements (Silhouettenkoeffizient, 2018).

$$S(o) = \frac{dist(B, o) - dist(A, o)}{max(dist(A, o), dist(B, o))} \tag{3.3}$$

Als Distanz zu einem Cluster gilt dabei die durchschnittliche Distanz des Elements o zu allen Objekten im Cluster. Den Silhouettenkoeffizienten des gesamten Clusterings bildet der Mittelwert aus allen Koeffizienten der einzelnen Elemente. Ein weiterer Indikator für die Qualität eines Clustering-Ergebnisses ist der erreichte Wert der RSS. Abbildung 3.8 zeigt die Silhouettenkoeffizienten und die negativen Werte der RSS zu Cluster-Anzahlen von zwei bis zwanzig. Die obere Kurve zeigt dabei die Silhouettenkoeffizienten und die untere die Werte der RSS.

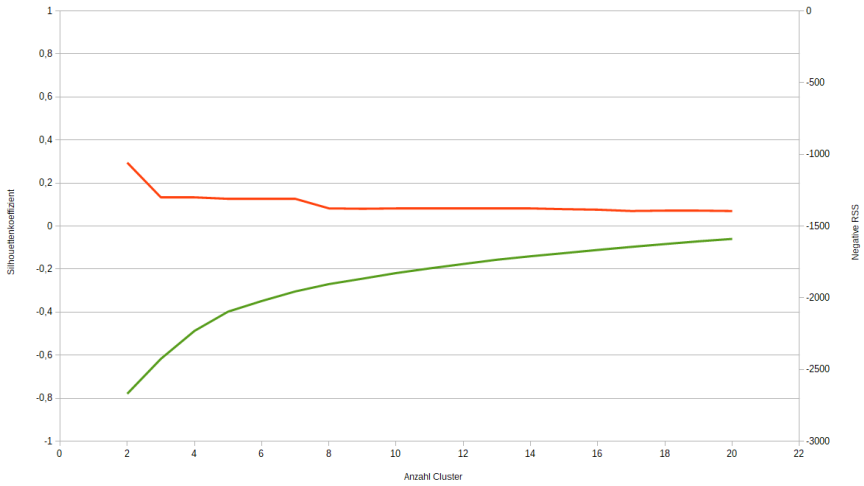


Abbildung 3.8.: Cluster-Scores bei verschiedenen Cluster-Anzahlen

Es ist erkennbar, dass der Silhouettenkoeffizient stetig sinkt während die negative RSS stetig steigt. Letzteres ist ein typisches Verhalten des K-means-Algorithmus. Der Wert nähert sich mit wachsender Cluster-Anzahl an null an. Stimmen die Anzahl der Datensätze und die vorgegebene Cluster-Anzahl überein, erreicht der Wert null (Schütze u. a., 2008, S. 365). Somit führt eine alleinige, naive Betrachtung dieses Wertes zu dem Schluss, dass eine Erhöhung der Cluster-

Anzahl immer auch zu einem besseren Ergebnis führt. Eine Möglichkeit, diesen Fehlschluss zu verhindern, ist es, einen Punkt zu ermitteln, an dem eine Erhöhung der Cluster-Anzahl um eins zu einer geringeren Verbesserung führt als die vorige. Bezogen auf eine grafische Darstellung der Werte wie in Abbildung 3.8 bedeutet dies, Knicke in der Kurve der Zielwerte zu suchen. In diesem Fall ist ein solcher Knick bei einer Cluster-Anzahl von fünf zu erahnen.

Der Silhouettenkoeffizient ist bei einer Cluster-Anzahl von zwei am höchsten. Diese ist in diesem Anwendungsfall allerdings nicht ausreichend und wird daher nicht weiter betrachtet. Für Cluster-Anzahlen von drei bis sieben bleibt der Koeffizient relativ Konstant. Daher folgt der Schluss, dass das bestmögliche Clustering für Cluster-Anzahlen in diesem Bereich liegen muss. In Kombination mit der Betrachtung der Zielwerte folgt der Schluss, dass eine Cluster-Anzahl von fünf für ein K-means-Clustering der Embeddings am besten geeignet ist.

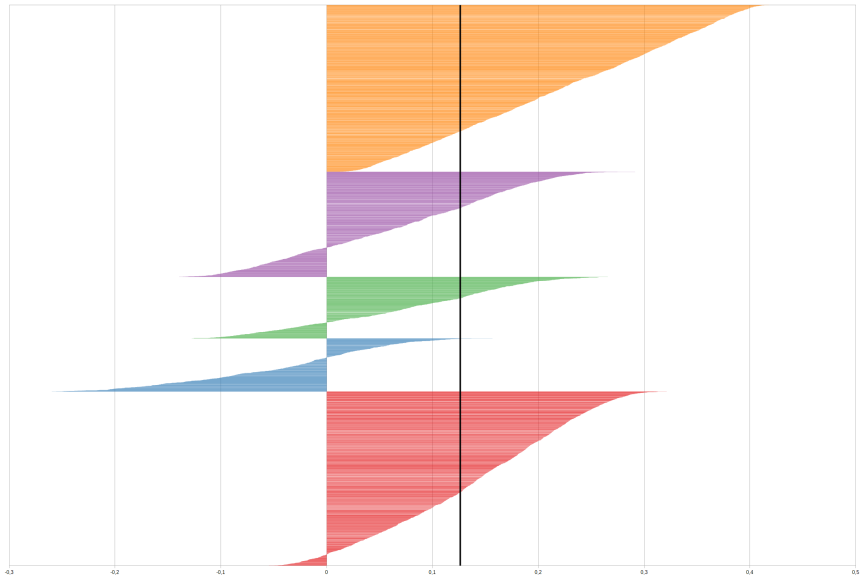


Abbildung 3.9.: Silhouettenplot der Cluster nach K-means-Clustering mit einer Cluster-Anzahl von fünf

Abbildung 3.9 zeigt den Silhouettenplot für das K-means-Clustering mit einer vorgegebenen Cluster-Anzahl von fünf. Diese Grafik zeigt zu jedem Datensatz die Silhouette nach Cluster und Silhouetten-Wert sortiert an. Die Schwarze Linie zeigt den Mittelwert der Silhouetten, also den Silhouettenkoeffizienten an. Die Einfärbung deckt sich dabei nicht mit Abbildung 3.7. Es ist erkennbar, dass die Cluster-Größen stark variieren. Dies ist aufgrund der Art der Daten

allerdings nicht zwingend ein Hinweis auf ein schlechtes Clustering. Es lässt sich ebenfalls sagen, dass in jedem Cluster Datensätze mit überdurchschnittlichen Silhouetten vorkommen. Insgesamt deutet diese Darstellung aber auf kein gutes Clustering hin. Diese Beobachtung wird dadurch unterstützt, dass der Silhouettenkoeffizient für sämtliche betrachtete Cluster-Anzahlen in einem eher negativ zu bewertenden Bereich liegt. Eine Betrachtung der Silhouettenplots der weiteren Werte von drei bis sieben weist keine besseren Ergebnisse auf. Daher bleibt die Festlegung der Cluster-Anzahl auf fünf bestehen. Abbildung 3.10 zeigt die Aufteilung des Vokabulars nach einem K-means-Clustering mit einer Anzahl von fünf Clustern. Die Cluster sind im Vergleich zu Abbildung 3.7 klarer abgegrenzt.

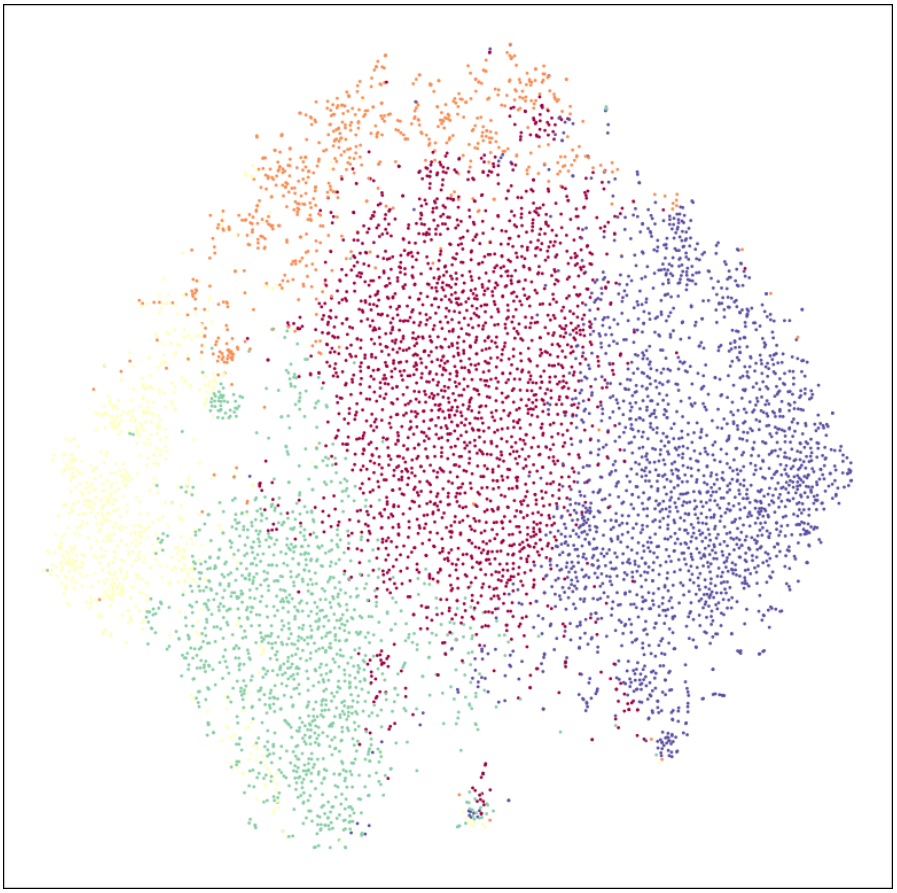


Abbildung 3.10.: Embeddings des gesamten Vokabulars nach Clustering mit einer Cluster-Anzahl von fünf

Alternativen zum K-means-Algorithmus

Eine mögliche Alternative zum K-means-Algorithmus ist *DBSCAN* (Ester u. a., 1996). DBSCAN ermittelt Cluster anhand der Dichte der Datensätze. Der Algorithmus erwartet ein Datenset, in dem sehr dichte Bereiche von Datensätzen durch weniger dichte Bereiche voneinander getrennt sind. Die Cluster-Bildung erfolgt dabei durch die Betrachtung der Umgebung jedes Elements im Datensatz. Befinden sich ausreichend viele Elemente in einem bestimmten Umkreis um ein Element, führt dies zur Bildung eines Clusters. Gegebenenfalls können nahe genug beieinander liegende Cluster zu einem vereint werden. Elemente, in deren Umgebung keine ausreichende Anzahl an anderen Elementen liegt, werden als Rauschen markiert. So entsteht ein Set aus Clustern und Rauschen. Dabei kann die Form der einzelnen Cluster beliebig sein. K-means neigt im Vergleich dazu, immer konvexe Cluster zu bilden. Diese Eigenschaft macht den Silhouettenkoeffizienten für Ergebnisse des DBSCAN-Algorithmus wenig geeignet, da dieser konvexe Cluster höher bewertet. Bei der Anwendung von DBSCAN sind die Größe des zu betrachtenden Bereichs um ein Element herum und die nötige Mindestanzahl an vorhandenen anderen Elementen in diesem Bereich zur Bildung eines Clusters setzbar. Die Größe des betrachteten Bereichs um ein Element herum wird mit *Epsilon* bezeichnet. An dieser Stelle findet die Implementierung aus der Scikit-Learn-Bibliothek Anwendung (Pedregosa u. a., 2011). Diese verwendet standardmäßig ein Epsilon von 0.5 und eine Mindestanzahl von 5.

Eine Anwendung des DBSCAN-Algorithmus in Standardparametrisierung ermittelt vier Cluster. Dabei entfallen allerdings 84,1% auf einen Cluster und 15,6% der Daten werden als Rauschen erkannt. Eine Verringerung des Epsilon auf 0.45 führt zur Bildung von acht Clustern. Dabei entfallen allerdings weiterhin 76,8% der Daten auf einen Cluster und 22,5% der Daten werden als Rauschen erkannt. Generell gilt hier die Beobachtung, dass eine Verringerung des Epsilon zur Bildung einer größeren Anzahl an Clustern führt. Dabei steigt allerdings die Menge an Rauschen mit sinkendem Epsilon. Eine Erhöhung des Epsilon führt zu gegenteiligem Verhalten. Generell bildet DBSCAN auf den Daten immer einen großen Cluster und gegebenenfalls mehrere sehr kleine, welche unter einem Prozent des Datensatzes ausmachen. Eine Erhöhung der Mindestanzahl verhindert die Bildung dieser kleinen Cluster. Dies führt allerdings wiederum dazu, dass mehr Elemente als Rauschen erkannt werden. Werden mehr als ein Cluster erkannt, bestehen die übrigen Cluster in der Regel weiterhin nur aus wenigen Elementen. Diese Beobachtungen führen zu dem Schluss, dass das Datenset nicht die erforderliche Dichtestruktur für ein Clustering mit DBSCAN aufweist.

Eingehende Betrachtung der Cluster

Nach der Ermittlung des am besten geeigneten Clusterings der Daten kann eine tiefere Betrachtung der ermittelten Cluster erfolgen. Diese bietet weitere Einblicke in die Struktur der Daten. Tabelle 3.3 zeigt die unterschiedlichen Cluster-Größen und die jeweils häufigsten Wörter innerhalb der Cluster nach einem Clustering der Word Embeddings mit K-means. Dabei sind die Ergebnisse ohne und mit vorangehender Normalisierung der Feature Vektoren gegenübergestellt.

Cluster	Vor Normalisierung		Nach Normalisierung	
	Größe	Häufigste Wörter	Größe	Häufigste Wörter
1	304	whiski, whiskey, one, bottl, bourbon, age, matur, dram, relea, glass	2777	old, includ, smell, reminisc, drop, straight, hand, fire, copper, sea
2	7550	like, come, open, appear, mayb, longer_averag, combin, influenc, fill, wine	3534	whiski, whiskey, one, bottl, bourbon, age, matur, averag, relea, glass
3	448	sweet, note, dri, vanilla, fruit, hint, spice, light, honey, fruiti	2634	dram, great, enjoy, seem, interest, offer, over_impress, certain, excel, power_oili
4	1810	rye, old, make, get, grain, averag, great, way, expect, someth	576	sweet, note, nose, dri, spice, slight, light, finish, oak, palat
5	354	nose, slight, finish, oak, palat, smoke, spici, quit, littl, tast	945	vanilla, fruit, hint, honey, floral, fresh, citrus, touch, toff, orang

Tabelle 3.3.: Vergleich der Cluster vor und nach Normalisierung der Embeddings

Die Tabelle unterstützt noch einmal die Beobachtung, dass die Cluster-Größen ausgeglichener sind. Ohne vorangehende Normalisierung nimmt einer der Cluster allein etwa 75% des ganzen Vokabulars ein. Ein Vergleich der häufigsten Wörter zeigt einige Übereinstimmungen und Unterschiede zwischen den Aufteilungen. Beide Aufteilungen weisen Cluster für Geschmacksbegriffe auf. Zudem befinden sich in beiden Varianten Cluster mit allgemeinen Begriffen aus dem Whisky-Bereich. Dies lässt sich darauf zurückführen, dass diese Wörter oft grammatikalisch eine ähnliche Rolle einnehmen.

Ein Blick auf die Häufigkeiten der einzelnen Wörter in den Clustern bietet weitere Erkenntnisse. Im Clustering ohne Vektornormalisierung befinden sich sämtliche Wörter, welche lediglich fünf oder sechs mal im Korpus vorkommen, im größten Cluster. Dies stützt die These, dass dieser Cluster vorwiegend irrelevante Wörter enthält. Die Cluster 1, 3 und 5 in dieser Variante beinhalten dagegen ausschließlich Wörter, welche mindestens 35 mal vorkommen. Bei den Clustern 3 und 5 liegt diese Zahl noch deutlich höher. Im zweiten Clustering befinden sich dagegen in jedem Cluster Wörter, welche nur fünf mal im Korpus vorkommen. Diese Beobachtungen unterstützen den Eindruck, dass das erste Clustering eher eine Unterteilung nach Relevanz und das zweite eher eine Unterteilung nach Bedeutung darstellt.

Die Tabelle und eine eingehendere Betrachtung der Cluster deuten des Weiteren auf einige Schwachstellen und Optimierungsmöglichkeiten des bisherigen Prozesses. Durch die doppelte Phrasenerkennung sind beispielsweise Kombinationen wie *glass classic malt colour* entstanden. Diese Kombination kommt im Korpus 603 mal vor. Es liegt nahe, dass dies das Resultat einer festen, wiederverwendeten Form für Tasting Notes bei einer der Textressourcen ist. Um den Einfluss solcher Schablonen zu verringern, bietet es sich an, Phrasen, welche aus mehr als drei Wörtern bestehen, aus dem Korpus zu entfernen. Zudem befinden sich die Begriffe *longer averag* mit 732 und *power oili* mit 641 Vorkommnissen unter den häufigsten Phrasen im Korpus. Dies gibt Anlass zu einer genaueren Betrachtung der Rohdaten. Diese ergibt, dass die beiden letztgenannten Phrasen aus der Ressource Lardin (2018) stammen. Daher steht an dieser Stelle der Beschluss, die Texte aus dieser Quelle aufgrund der scheinbar mangelhaften Qualität für die weitere Arbeit nicht mehr zu berücksichtigen. Die Phrase *glass classic malt colour* lässt sich auf einen bestimmten Autor der Ressource Whisky Monitor (2018) zurückführen. Da dies wie bereits erwähnt auf eine Art Textschablone hindeutet und lange Phrasen leicht zu entfernen sind, können die Texte dieses Autoren zunächst beibehalten werden.

Zwischenfazit

Diese Erkenntnisse führen zu einem Rückschritt im KDD-Prozess. Die Entfernung der Ressource Lardin (2018) stellt eine Rückkehr zur Datenauswahl dar. Zudem muss eine Ermittlung von Phrasen, welche aus mehr als drei Wörtern bestehen, erfolgen. Dies stellt eine Optimierung der Datenvorverarbeitung dar. Die weiteren Schritte dazwischen bleiben unverändert. Zusätzlich zu den beiden genannten Änderungen erfolgt eine Anpassung der Parametrisierung des Word2Vec-Algorithmus bei der anschließenden Generierung der Word Embeddings. Wörter,

welche seltener als zehn mal vorkommen, sollen nicht mehr berücksichtigt werden. In der Standardparametrisierung des Algorithmus liegt dieser Wert bei fünf.

3.7. Fünfter Durchlauf der Datenaufbereitung

Nach den Erkenntnissen aus dem vorigen Durchlauf muss ein Rückschritt zur Datenauswahl erfolgen. Diese muss insofern korrigiert werden, dass die Ressource Lardin (2018) aufgrund mangelnder Qualität ausgeschlossen wird. Des weiteren soll das Preprocessing um die Entfernung zu langer Phrasen erweitert werden und die Mindestanzahl an Vorkommnissen eines Wortes, um beim Training berücksichtigt zu werden, auf zehn gesetzt werden.

3.7.1. Erweiterung des Preprocessings um die Entfernung zu langer Phrasen

Im Anschluss an die Phrasenerkennung sollen solche, welche sich aus mehr als drei Wörtern zusammensetzen, entfernt werden. Die am einfachsten umzusetzende Möglichkeit dazu ist eine Iteration über sämtliche Texte im Korpus, bei der diese Phrasen entsprechend ermittelt und entfernt werden. Listing 3.6 zeigt ein Codebeispiel, wie dies in Python umsetzbar ist.

```
1 for text in texts:
2     text = [[word for word in sentence if len(word.split('_')) < 4]
3             for sentence in text]
```

Listing 3.6: Beispielcode zur Erkennung und Entfernung zu langer Phrasen im Korpus

Der Code geht dabei davon aus, dass der Korpus bereits in Form von verschachtelten Listen vorliegt und dass als Verbindungszeichen für Phrasen Unterstriche verwendet wurden.

3.7.2. Generierung der Word Embeddings

Im Anschluss an die Entfernung der Phrasen erfolgt ein erneutes Training mit dem Word2Vec-Algorithmus. Dabei ist die Mindestanzahl an Vorkommnissen pro Wort, um berücksichtigt zu werden, auf 10 gesetzt. Dies hat direkt zur Folge, dass die Größe des Embedding-Vokabulars von über 10.000 auf etwa 7.000 Wörter reduziert wird.

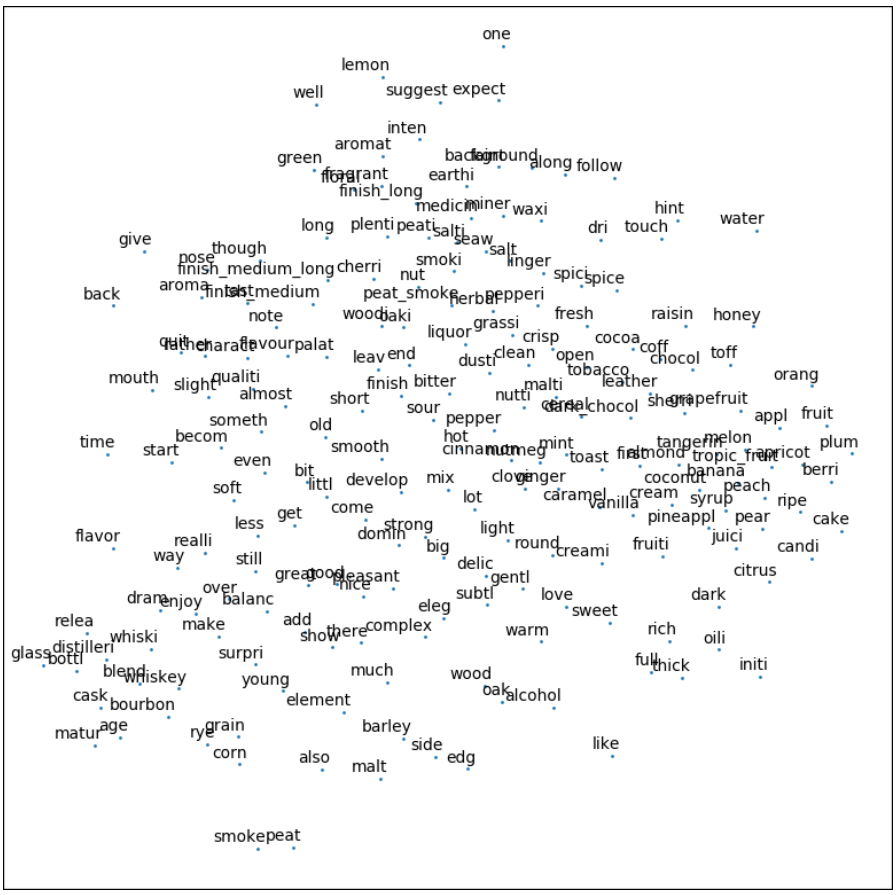


Abbildung 3.11.: Embeddings der 200 häufigsten Wörter im Datensatz nach Entfernung zu langer Phrasen

Abbildung 3.11 zeigt die 200 häufigsten Wörter nach Anwendung der genannten Änderungen. Ein nennenswerter Unterschied zu Abbildung 3.4 ist auf den ersten Blick nicht zu erkennen. Da es sich hierbei um die häufigsten Wörter handelt, ist dies auch nicht zu erwarten.

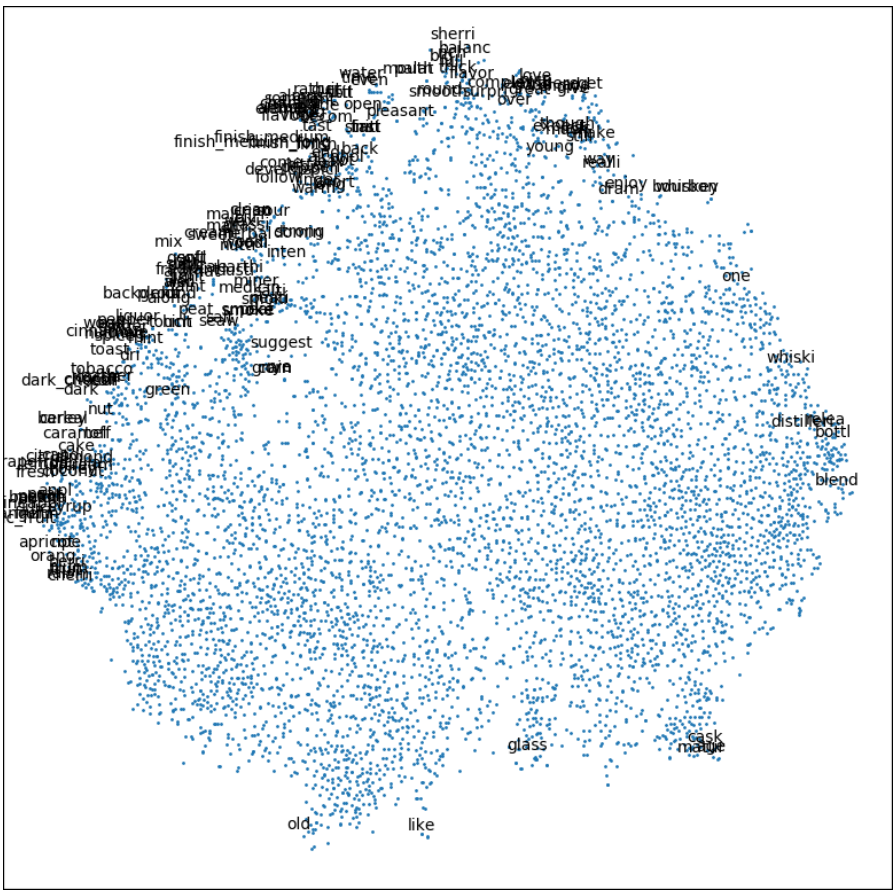


Abbildung 3.12.: Embeddings des gesamten Vokabulars nach Entfernung zu langer Phrasen

Abbildung 3.12 zeigt alle Wörter des Datensets. Die 200 häufigsten Wörter sind dabei gelabelt. Auch hier ist kein nennenswerter Unterschied zu Abbildung 3.5 zu erkennen.

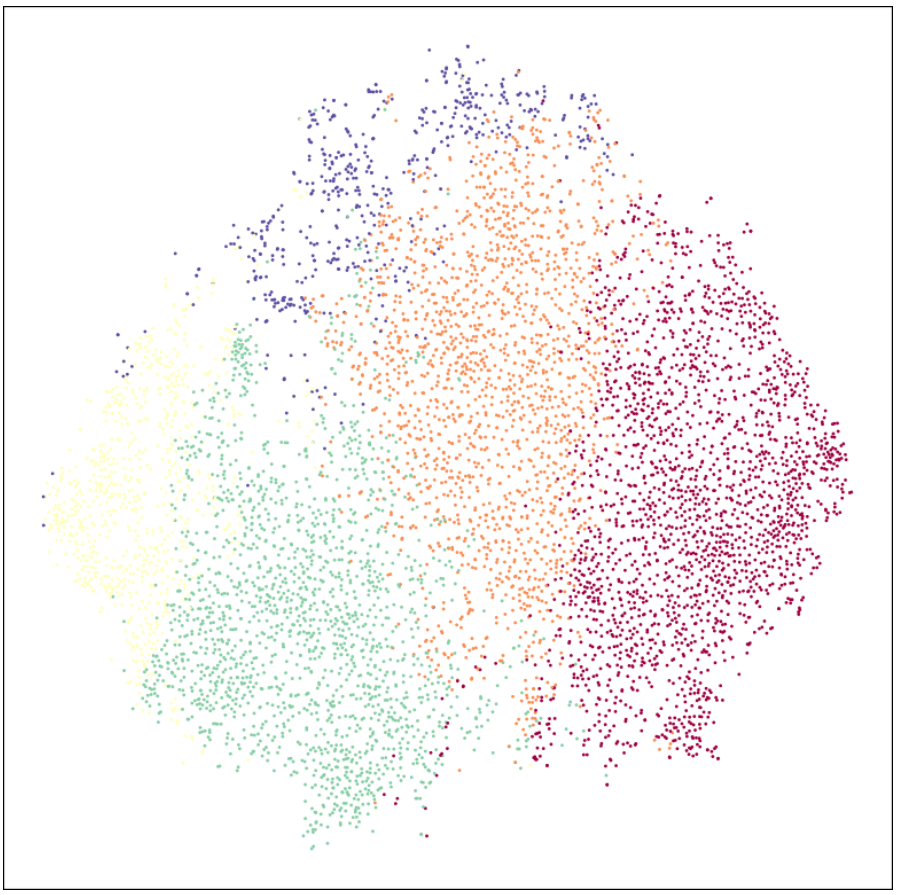


Abbildung 3.13.: Embeddings des gesamten Vokabulars nach Entfernung zu langer Phrasen, eingefärbt nach Clustern

Abbildung 3.13 zeigt die Daten nach dem anschließenden Clustering anhand der normalisierten Vektoren. Dieses erscheint ebenso sehr nah an Abbildung 3.7 zu sein. Um einen Einfluss der angewandten Maßnahmen zu erkennen muss also eine Betrachtung der Cluster erfolgen.

Cluster	Größe	Häufigste Wörter
1	2110	whiski, whiskey, one, bottl, bourbon, age, matur, relea, glass, cask
2	1793	dram, enjoy, seem, interest, offer, over_impress, certain, perhap, that, without
3	818	fruit, hint, vanilla, light, honey, fresh, like, floral, touch, toff
4	1808	old, someth, iodin, reminisc, sulphur, cut, straight, fire, ear, sea
5	539	sweet, note, nose, dri, slight, spice, finish, oak, palat, fruiti

Tabelle 3.4.: Cluster nach Entfernung zu langer Phrasen

Tabelle 3.4 zeigt die Größen der Cluster und die jeweils häufigsten Wörter. Der Vergleich mit Tabelle 3.3 weist eine hohe Übereinstimmung bei den häufigsten Wörtern auf. Die Größenverhältnisse sind ebenfalls ähnlich geblieben. Unterschiede sind das fehlen der störenden Phrasen und allgemein das kleinere Vokabular. Anhand des Vergleichs lässt sich feststellen, dass sich die Qualität der Embeddings scheinbar nicht in großem Maße verändert hat. Dennoch wurde erreicht, dass einige unerwünschte Begriffe entfernt sind.

3.8. Weitere Optimierungen des Wortvektortrainings

Der Word2Vec-Algorithmus bietet weitere Möglichkeiten zur Parametrisierung. Diese bergen das Potential, die Embeddings weiter zu verbessern. Ein wichtiger Faktor ist die Auswahl des Kernalgorithmus. Standardmäßig verwendet die Word2Vec-Implementierung CBOW. Die Alternative Skip-Gram soll getestet werden. Diese soll bei kleineren Datensets bessere Ergebnisse liefern (Mikolov, 2013).

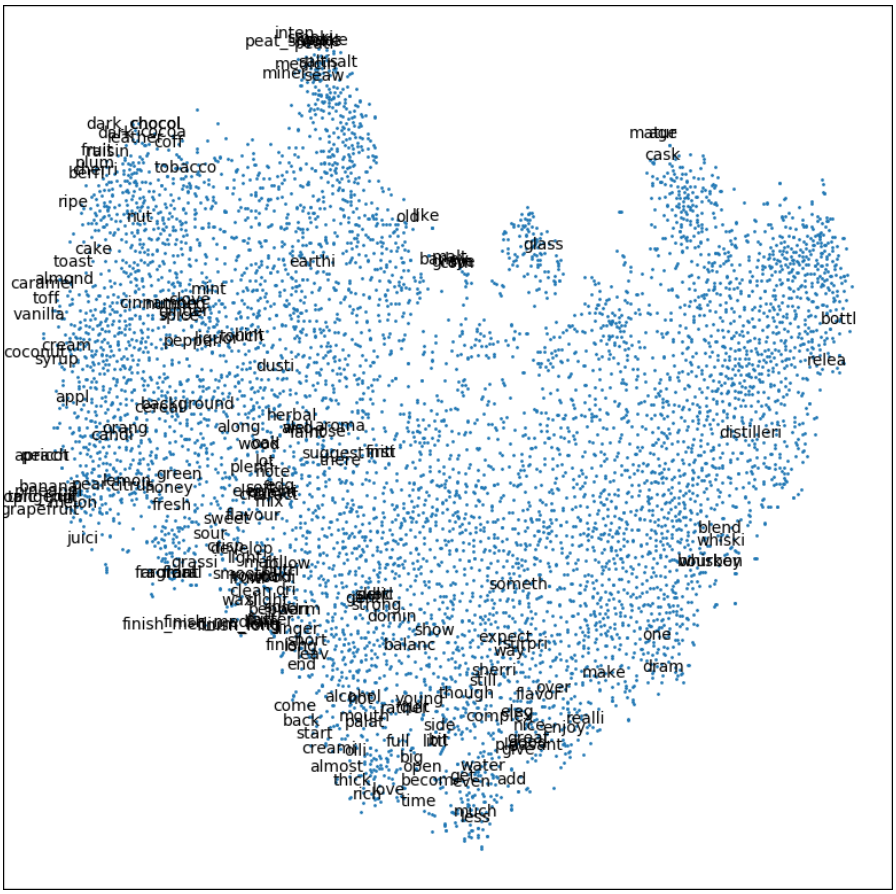


Abbildung 3.14.: Embeddings des gesamten Vokabulars, generiert mit Skip-Gram

Abbildung 3.14 zeigt das gesamte Datenset nach einem Training der Word Embeddings mit Skip-Gram als Basisalgorithmus. Die 200 häufigsten Wörter sind gelabelt. Im Vergleich zu vorigen Varianten liegen diese nicht mehr ausschließlich am Rand der Grafik. Dies spricht für

einen geringeren Einfluss der Häufigkeiten der Wörter auf die Embeddings und somit für eine bessere Abbildung der Bedeutungen.

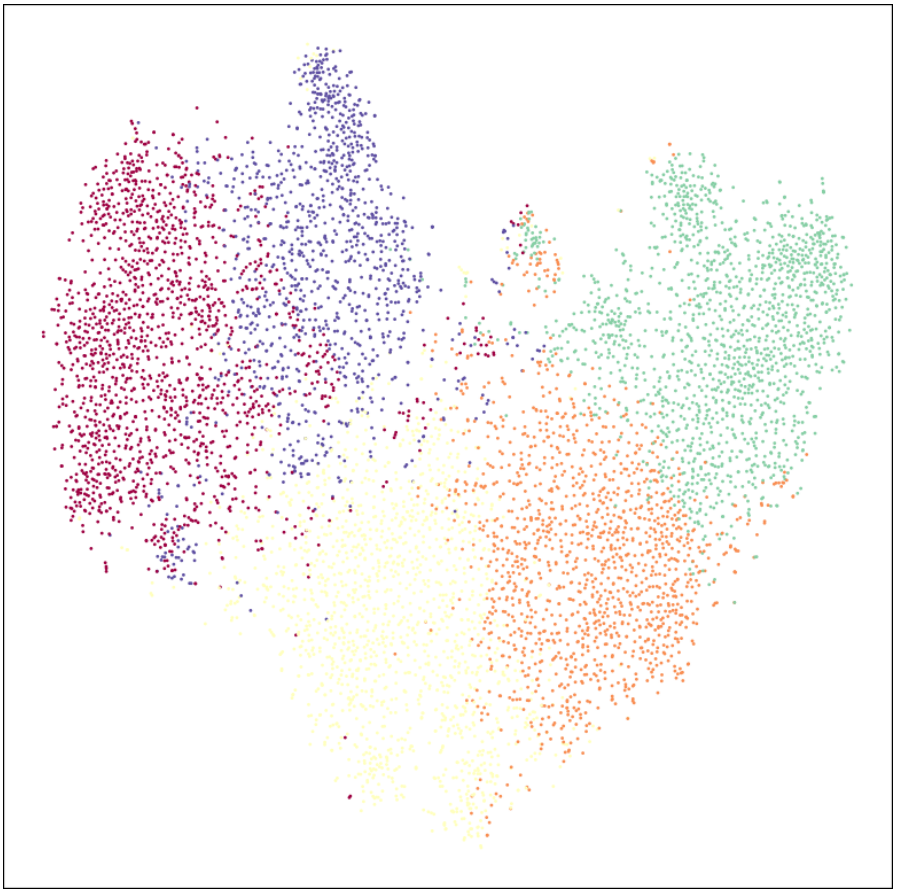


Abbildung 3.15.: Embeddings des gesamten Vokabulars nach Clustering, generiert mit Skip-Gram

Abbildung 3.15 zeigt das gleiche Datenset nach einem Clustering. Rein optisch lässt sich kein großer Unterschied an der Qualität des Clusterings im Vergleich zu vorigen Grafiken ausmachen. Es scheint allerdings, als seien die Cluster in ihrer Größe ausgeglichener. Dies belegt ein tieferer Blick auf die Cluster.

Cluster	Größe	Häufigste Wörter
1	1583	fruit, hint, vanilla, honey, fresh, floral, touch, toff, orang, lot
2	1360	whiski, whiskey, one, much, bourbon, age, make, matur, dram, glass
3	1509	sweet, nose, dri, slight, light, spice, finish, oak, palat, fruiti
4	1525	bottl, relea, cask, blend, distilleri, express, new, barrel, distil, color
5	1091	note, smoke, like, well, wood, smoki, leather, peati, earthi, peat_smoke

Tabelle 3.5.: Cluster nach Training unter Verwendung von Skip-Gram

Tabelle 3.5 zeigt die Cluster-Größen und die jeweils häufigsten Wörter der Cluster. Es ist erkennbar, dass das Verhältnis der Cluster-Größen deutlich ausgeglichener ist als bei vorigen Versuchen. Die Verteilung der häufigsten Wörter ist ebenfalls deutlich verbessert. So sind Cluster für fruchtig-süße und rauchig-erdige Aromen erkennbar. Dazu scheint ein Cluster Hauptgeschmacksmerkmale zu vereinen. Es verbleiben zwei Cluster mit beschreibenden Begriffen aus der Domäne Whisky. Dies war bereits in vorigen Aufteilungen der Fall.

3.8.1. Negative Sampling

Eine Methode, den Einfluss häufig vorkommender Wörter zu verringern ist das *Negative Sampling* (Mikolov u. a., 2013b). Während des Trainings eines Neuronalen Netzwerks werden bei der Verarbeitung jedes Datensatzes alle Gewichtungen in der Gewichtungsmatrix angepasst. Dies führt im Falle von Word Embeddings dazu, dass besonders häufige Wörter den meisten Einfluss auf die Matrix haben. Negative Sampling beschreibt die Methode, bei jeder Verarbeitung eines Datensatzes lediglich die Gewichtungen einer fixen Anzahl zufällig ausgewählter anderer Wörter anzupassen. Dabei haben häufigere Wörter eine höhere Wahrscheinlichkeit, als Negative ausgewählt zu werden. Dies führt dazu, dass sich die häufigeren Wörter gegenseitig mehr beeinflussen als die selteneren. Mikolov u. a. (2013b) nennt für kleinere Textkorpora einen Wert von fünf bis zwanzig für die Anzahl der zu wählenden Negative. In der Standardparametrisierung verwendet Word2Vec fünf Negative. Im Folgenden soll der Effekt einer Anpassung dieses Wertes auf zwanzig betrachtet werden.

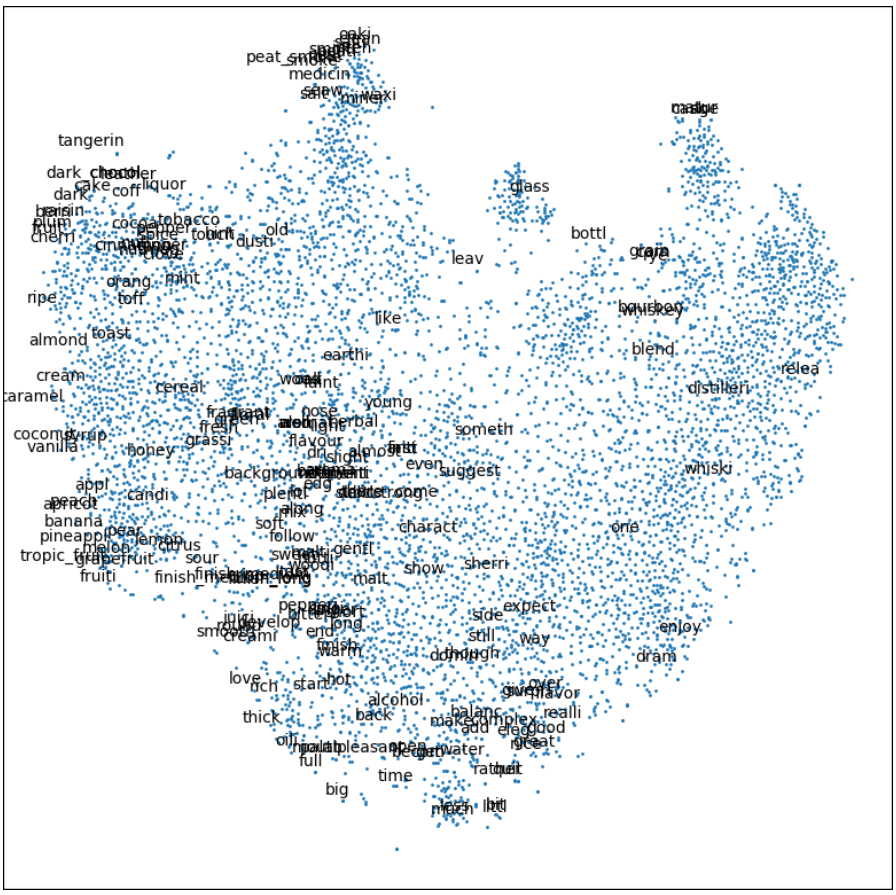


Abbildung 3.16.: Word Embeddings nach Erhöhung des Negative Samplings

Abbildung 3.16 zeigt die Verteilung der Wörter nach Anpassung des Negative Samplings. Die häufigsten 200 Wörter sind gelabelt. Im Vergleich zu Abbildung 3.14 ist erkennbar, dass die häufigsten Wörter noch einmal etwas gleichmäßiger über das Datenset verteilt sind. Dies ist mit der Verringerung des Einflusses dieser auf die Gewichtungsmatrix zu erklären.

Cluster	Größe	Häufigste Wörter
1	1527	fruit, hint, vanilla, honey, fresh, floral, toff, orang, lot, caramel
2	1561	sweet, nose, dri, slight, light, spice, finish, oak, palat, fruiti
3	1367	whiski, good, whiskey, one, realli, much, bourbon, make, dram, glass
4	1472	bottl, age, matur, relea, cask, blend, distilleri, offer, express, new
5	1141	note, smoke, touch, well, wood, smoki, peati, earthi, peat_smoke, liquor

Tabelle 3.6.: Cluster nach Erhöhung des Negative Samplings

Tabelle 3.6 zeigt die Cluster-Größen und die häufigsten Wörter nach Erhöhung des Negative Sampling. Das Größenverhältnis ist ähnlich zum vorigen Clustering. Anhand der häufigsten Wörter lässt sich kein nennenswerter Unterschied ausmachen. Dennoch scheint nach Betrachtung der Grafik 3.16 eine Verbesserung vorzuliegen.

3.8.2. Anpassung der Kontextgröße

Eine weitere Möglichkeit, die Generierung der Word Embeddings anzupassen, ist eine Veränderung der Kontextgröße. Diese ist in der Standardparametrisierung auf fünf festgelegt. Das bedeutet, dass während der Initialisierung des Trainings zu jedem Wort Paare mit allen Wörtern in einem Umkreis von fünf Wörtern innerhalb desselben Satzes gebildet werden. Eine Erhöhung des Kontextes kann möglicherweise zu einer Verbesserung der Embeddings führen.

Abbildung 3.17 zeigt die Cluster der Embeddings nach Vergrößerung des Kontextes. Eine interessante Beobachtung ist, dass die Cluster nun eher die Form von Balken annehmen. Die häufigsten Wörter scheinen erneut noch weiter über das Datenset verteilt zu sein. Eine Betrachtung der Aufteilung der Wörter auf die Cluster stellt erneut keine großen Veränderungen heraus.

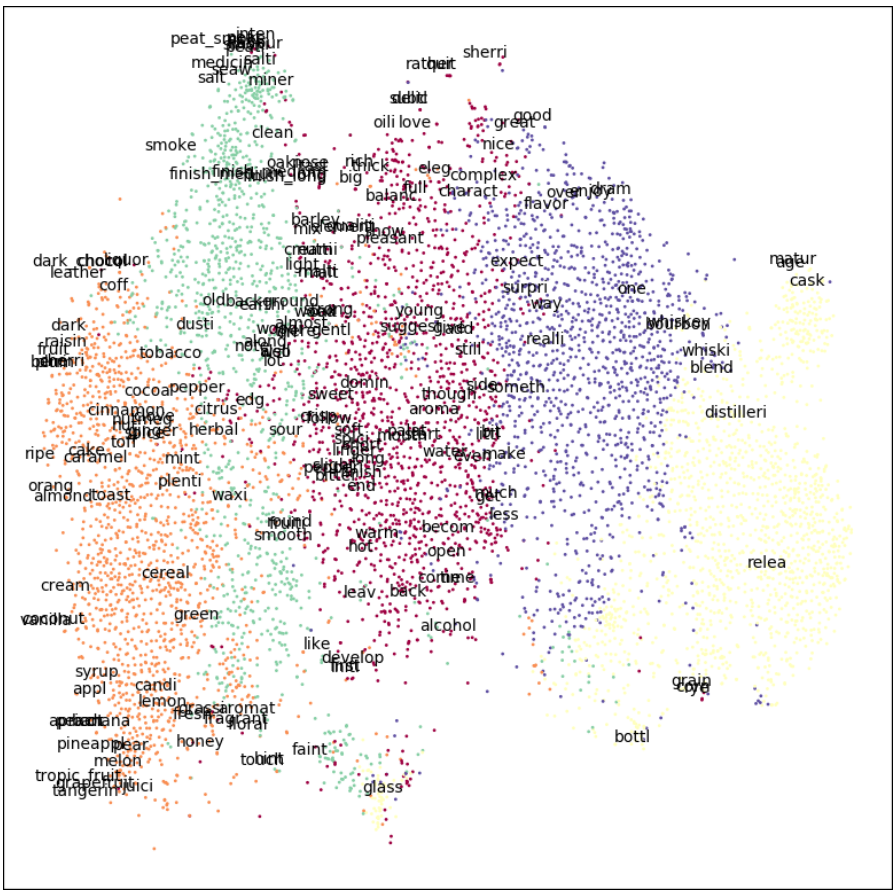


Abbildung 3.17.: Cluster der Word Embeddings nach Vergrößerung des Kontexts

3.8.3. Anpassung der Vektorgröße

Zuletzt ist auch die Größe der resultierenden Embeddings anpassbar. In der Standardparametrisierung beträgt dieser Wert 100. Eine Erhöhung dieses Wertes führt dazu, dass mehr Neuronen im Hidden Layer die Gewichtungsmatrix bilden. Generell bedeutet ein höherer Wert damit eine feinere Berechnung der Wahrscheinlichkeiten. Gleichzeitig führt dies allerdings auch zu einer höheren Rechenlast. Eine Erhöhung muss nicht immer zu besseren Ergebnissen führen. Im Folgenden soll der Einfluss dieses Parameters durch eine Verdopplung der Vektorgröße Betrachtet werden.

Abbildung 3.18 zeigt alle Daten und Cluster nach Erhöhung der Vektorgröße auf 200. Es sind auf den ersten Blick erneut keine wesentlichen Unterschiede zu erkennen. Ebenso finden

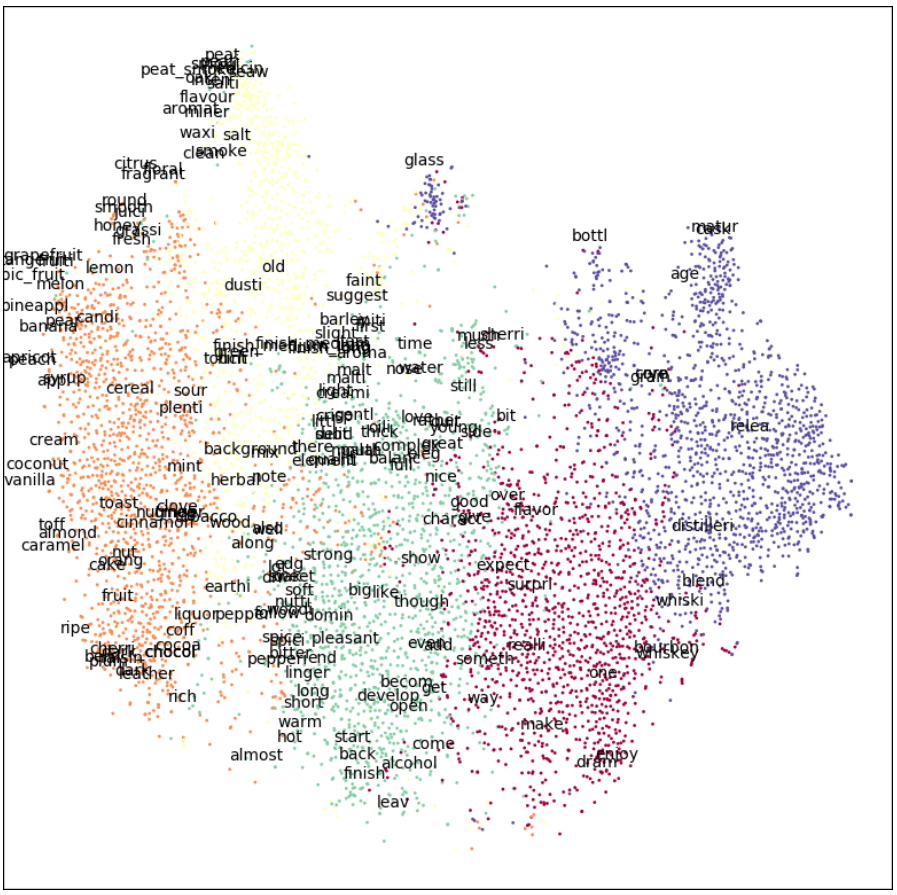


Abbildung 3.18.: Cluster der Word Embeddings nach Vergrößerung der Vektoren

sich bei einer Betrachtung der Cluster-Aufteilung keine nennenswerten Unterschiede. Dies spricht dafür, dass die Optimierung der Generierung der Word Embeddings an diesem Punkt weitestgehend ausgereizt ist.

3.8.4. Weitere Optimierungsmöglichkeiten

Bei einer Betrachtung des Vokabulars fallen weiterhin einige Unstimmigkeiten auf. So befinden sich darin auch nach der Entfernung aus vier Wörtern bestehender Begriffe nicht sinnvolle Wortkombinationen. Dies zeigt ein Blick auf die nächstgelegenen Begriffe zum Wort *mango*.

Tabelle 3.7 zeigt die zwanzig nächstgelegenen Begriffe zu *mango*. In dieser Auflistung befinden sich allein elf sinnlose Wortkombinationen. Diese sind allerdings eher selten. Daher

Begriff	Häufigkeit	Begriff	Häufigkeit
mango_passion_fruit	29	papaya_mango	25
guava	148	mango_papaya	47
passion_fruit	332	pineappl_papaya	20
tin_pineappl	78	guava_mango	20
mango_guava	25	mango_pink_grapefruit	11
papaya	185	litchi	11
guava_papaya	13	mango_pineappl	32
pineappl_mango	45	whitecurr	33
maracuya	12	passion_fruit_mango	17
yellow_plum	100	quinc_jam	11

Tabelle 3.7.: Nächstgelegene Begriffe zu mango

soll untersucht werden, wie sich eine Erhöhung der Mindestanzahl an Vorkommnissen eines Wortpaares, um als zusammenzufassender Begriff in Betracht gezogen zu werden, auf fünfzig auswirkt. Dies beinhaltet auch, dass einige sinnvolle Kombinationen wie *quinc_jam* verloren gehen. Diese Änderung bedeutet einen erneuten Rückschritt zur Preprocessing-Phase.

Begriff	Häufigkeit	Begriff	Häufigkeit
papaya	341	guava	253
passion_fruit	411	maracuya	12
tin_pineappl	78	kumquat	130
litchi	11	kiwi	292
ripe_banana	182	whitecurr	33
smoothi	17	persimmon	21
honeydew_melon	146	physali	10
honeydew	21	cantaloup	44
plantain	19	maracuja	14
salad	55	lilt	10

Tabelle 3.8.: Nächstgelegene Begriffe zu mango nach Anpassung der Phrase Detection

Tabelle 3.8 zeigt die nächsten Nachbarn zu *mango* nach dieser Anpassung. Die genannten, unerwünschten Begriffe fehlen. Das Vokabular ist auf 5180 Wörter verkleinert.

Abbildung 3.19 zeigt die Embeddings nach der Anpassung der Phrase Detection. Hier sind keine nennenswerten Unterschiede zu vorigen Ergebnissen zu erkennen.

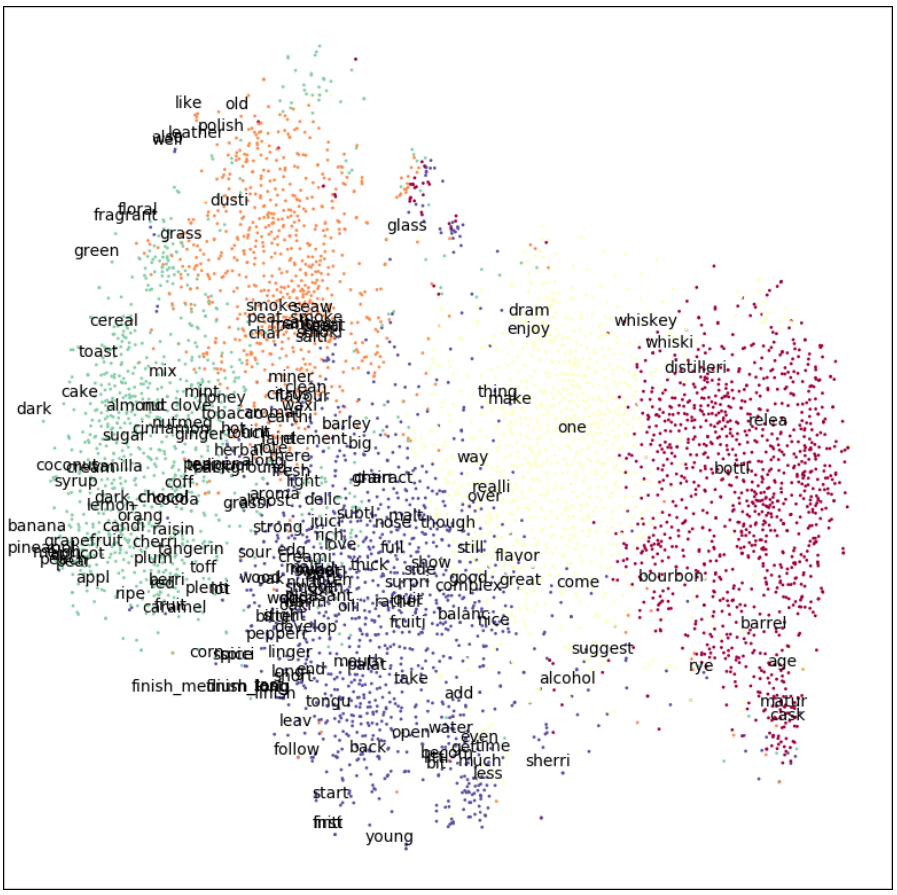


Abbildung 3.19.: Cluster der Word Embeddings nach Anpassung der Phrase Detection

3.9. Anwendung der Wortvektoren auf das Testdatenset

Nachdem die Generierung der Word Embeddings Optimiert ist, können diese zur Bildung der Repräsentationsform der Whiskys genutzt werden. Zu jedem Whisky im Testdatenset kann nun anhand der in den zugehörigen Tasting Notes vorkommenden Wörtern ein Set von Embeddings ermittelt werden. Um dies zu ermöglichen müssen die Tasting Notes des Testdatensets zunächst genau so vorverarbeitet werden wie die des Trainingsdatensets. Dies umfasst das Entfernen von Stop-Words, das Stemming und die Phrase Detection. Die Entfernung von Stop-Words ist dabei nicht zwingend notwendig, da zu diesen Wörtern keine Embeddings ermittelt wurden. Dennoch bietet dieses Vorgehen mindestens eine geringere Auslastung des Arbeitsspeichers. Durch eine Kombination der Embeddings eines Whiskys lässt sich ein Repräsentativer Vektor zu diesem

bilden. Hierfür wird der Mittelvektor aus seinem Vektorset verwendet. Dadurch entsteht ein Datenset aus Whisky-Vektoren, welches die gleichen Operationen wie in den vorangegangenen Schritten ermöglicht.

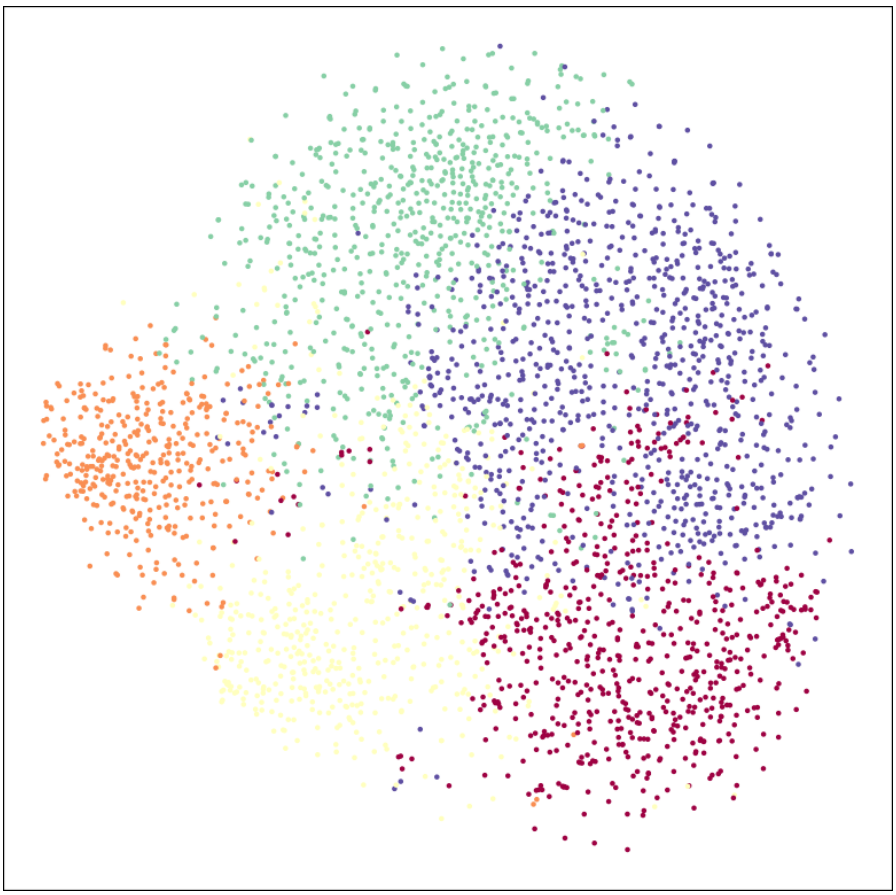


Abbildung 3.20.: Whiskys nach Vorverarbeitung, Anwendung der Word Embeddings und anschließendem Clustering

Abbildung 3.20 zeigt die Whisky-Vektoren nach Clustering. Es ist erkennbar, dass die Cluster im Vergleich zu den Wort-Clustern weniger klar abgegrenzt zueinander sind. Dies lässt sich damit erklären, dass das Set aus Whisky-Vektoren noch dichter ist. Tiefere Erkenntnisse verspricht eine Eingehende Betrachtung der Cluster. Ein Scoring der Cluster anhand einer *Ground Truth* ist für dieses Datenset aufgrund der in Kapitel 2.3.2 beschriebenen Komplexität der Whisky-Kategorien nicht sinnvoll, beziehungsweise aussagekräftig.

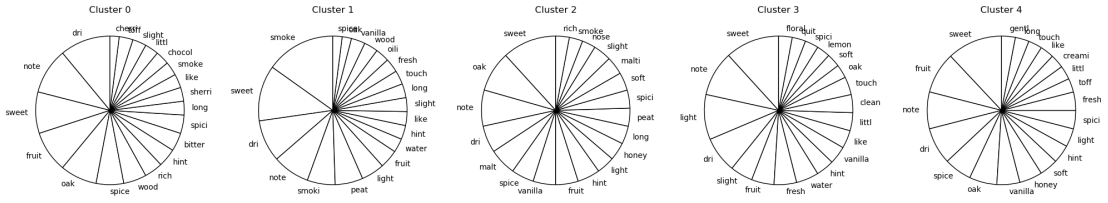


Abbildung 3.21.: Häufigste Wörter je Cluster in Relation zur Gesamtgröße aller Texte eines Clusters

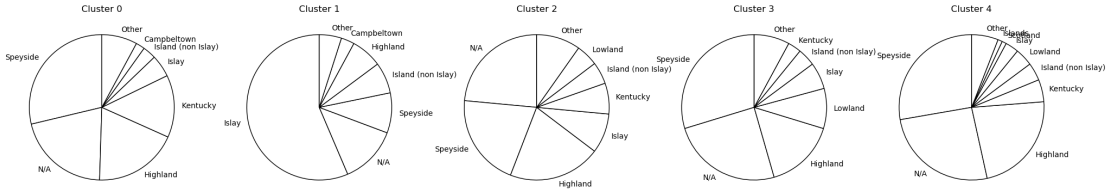


Abbildung 3.22.: Häufigste Regionen nach Clustern

Abbildung 3.21 zeigt die Häufigsten Wörter je Cluster. Dabei ist die Relation zur Gesamtanzahl an Wörtern innerhalb des jeweiligen Clusters berücksichtigt. Es ist erkennbar, dass die Wörter *sweet*, *note* und *dri* immer unter den häufigsten sind. Besonders *sweet* ist in drei Fällen das häufigste Wort. Dies liegt unter anderem daran, dass der Begriff der häufigste im Gesamten Datenset ist und gleichzeitig einen sehr allgemeinen Geschmack beschreibt. Auffällig ist der Cluster 1. In diesem dominieren als einzigem Cluster Begriffe für rauchige Noten wie *smoke*, *smoki* und *peat*. Dies deckt sich mit Abbildung 3.22. Die Grafik zeigt, dass dieser Cluster als einziger von Whiskys von der Insel Islay dominiert wird, während sich die Verteilungen in den anderen Clustern untereinander sehr ähneln. Islay-Whiskys sind für ihre besonders rauchigen Aromen bekannt. Daher ist diese Beobachtung als Erfolg zu bewerten. Dass die anderen Cluster derart starke Überschneidungen zeigen ist dagegen nicht notwendigerweise negativ zu bewerten. Die Herkunft der Whiskys dient hier lediglich als Vergleich. Eine weitere Beobachtung ist, dass das Testdatenset offensichtlich sehr stark von schottischen Whiskys dominiert wird.

Die an dieser Stelle ermittelten Whisky-Vektoren sollen als Grundlage für die Evaluierung der Arbeit dienen.

4. Evaluierung der Versuchsergebnisse

Um eine Aussage über die Qualität der generierten Whisky-Embeddings treffen zu können, muss eine Evaluierung dieser Stattfinden. Hierfür stehen verschiedene Methoden zur Verfügung. Eine Möglichkeit ist es, die ermittelten Cluster mit bestehenden Kategorien abzugleichen. Zudem gibt es Methoden, die die Qualität von Clustern in Form von Scores ermitteln. Außerdem besteht die Möglichkeit, die Qualität der Empfehlungen und Cluster von Experten bewerten zu lassen. Eine weitere Möglichkeit ist ein Abgleich der Empfehlungen mit *Flavour Maps*. Dies sind zweidimensionale Graphen, auf denen ähnliche Whiskys nahe beieinanderliegen.

4.1. Bewertung der verschiedenen Methoden

Ein Abgleich der ermittelten Cluster mit bestehenden Kategorien erscheint zunächst sinnvoll, ist aber wenig erfolgversprechend. Ein Vergleich allein mit einer Aufteilung nach Herkunft oder Alter ist aufgrund der Vielfalt allein innerhalb einzelner Länder nicht sinnvoll. Das Alter eines Whiskys hat je nach Herstellungsregion eine sehr unterschiedliche Auswirkung auf den Geschmack. Eher geeignete Kriterien für einen solchen Vergleich sind das verwendete Rohmaterial, der verwendete Fasstyp und der Alkoholgehalt.

Eine Möglichkeit, die Cluster mit echten Kategorien zu vergleichen, bieten verschiedene Scoring-Algorithmen. Diese vergleichen die Cluster mit einer *Ground Truth*. Ebenso existieren Algorithmen, die keine Ground Truth verwenden, sondern die Qualität der Cluster an sich bewerten. Die so ermittelten Werte treffen dadurch keine Aussage über die Qualität der Empfehlungen sondern lediglich über die Cluster im Verhältnis zu den Daten und zueinander. Da eine Ground Truth für Whisky-Kategorien aufgrund ihrer Komplexität nicht einfach erstellt werden kann, ergibt eine Untersuchung in diese Richtung keinen Sinn.

Ein Vergleich der Daten mit einer Flavour Map gestaltet sich als schwierig, da diese in der Regel nur wenige Whiskys darstellen.

Die Methode, die ermittelten Empfehlungen und Cluster von Experten bewerten zu lassen, verspricht die aussagekräftigsten Ergebnisse. Die Akzeptanz der Empfehlungen unter Experten

ist ein geeignetes Maß dafür, wie sehr ein System, welches auf ihnen basiert, als Expertensystem akzeptiert würde. Ein Problem bei einer Expertenbefragung ist die Verfügbarkeit von geeigneten Experten, welche bereit sind, eine solche Bewertung durchzuführen. Da Whisky ein sehr großes Feld ist und das verwendete Testdatenset entsprechend groß ist, ist es schwer, Personen zu finden, welche sich ausreichend mit Whisky auskennen. Daher muss an dieser Stelle eine relativ geringe Anzahl an Experten ausreichen.

4.2. Durchführung der Evaluierung

Nach der Bewertung der unterschiedlichen Möglichkeiten zur Evaluierung der Ergebnisse und der Entscheidung für die Befragung von Experten muss ein geeigneter Fragebogen für diese erstellt werden. Zur Durchführung einer Expertenbefragung eignet sich ein Fragebogen.

4.2.1. Konzipierung eines Fragebogens

Ziel des Fragebogens ist es, zu ermitteln, wie akzeptabel die Empfehlungen und Cluster sind. Dazu sollen einerseits Fragen gestellt werden, welche darauf abzielen, Empfehlungen bewerten zu lassen und andererseits Fragen, welche darauf abzielen, die Cluster bewerten zu lassen. Um die Empfehlungen möglichst vollständig bewerten zu können, sollen dem Befragten zu einem Referenz-Whiksy die nach Berechnung am nächsten liegenden Whiskys vorgeschlagen werden. Diese soll der Befragte dann anhand ihrer Nachvollziehbarkeit als Empfehlung bewerten. Hierfür muss ein repräsentatives Set an Whiskys für den Fragebogen ausgewählt werden. Dabei ist das Ziel, mit möglichst wenigen Whiskys die Vielfalt der Domäne abzubilden und gleichzeitig Whiskys auszuwählen, die bekannt sind. Dafür wurde aus den bestehenden Whisky-Kategorien jeweils einer der bekanntesten ausgewählt. Um die Umfrageteilnehmer nicht abzuschrecken, wurde die Anzahl der Whiskys auf acht reduziert. Die ausgewählten Whiskys sind *Lagavulin 26 Years Old*, *Glenmorangie 10 Years Old*, *The Balvenie 12 Years Old, Double Wood*, *Jameson Irish Whiskey*, *Elijah Craig 12 Years Old*, *Talisker 10 Years Old*, *Nikka From the Barrel* und *Auchentoshan Three Wood*. Zur Bewertung der Cluster sollen den Befragten vier Sets an Whiskys vorgestellt werden, welche er dann anhand ihrer Nachvollziehbarkeit als Zusammenstellung bewerten soll. Aufgrund der großen Datenmenge erweist sich dies als schwierig. Als Sets wurden hier in zwei Fällen eine zufällige Auswahl aus einem Cluster gewählt, einmal die am nächsten am Cluster-Center gelegenen Whiskys und einmal die am weitesten vom Cluster-Center entfernten Whiskys gewählt. Dabei stammt jedes Set aus einem anderen Cluster.

Der resultierende Fragebogen befindet sich in Kapitel B.1. Im Rahmen der Durchführung der Umfrage wird dieser als Ausdruck an qualifizierte Personen verteilt und zusätzlich öffentlich verfügbar im Internet bereitgestellt.

4.2.2. Auswertung der Umfrageergebnisse

Im Anschluss an die Durchführung der Umfrage muss eine Auswertung der ausgefüllten Fragebögen erfolgen. Insgesamt haben acht Experten teilgenommen. Davon nahmen vier Personen über den ausgedruckten Fragebogen teil. Hierbei handelt es sich um Verkäufer aus Wein- und Spirituosen-Geschäften. Die weiteren vier Teilnehmer nutzten den Online-Fragebogen. Von diesen gaben alle an, sich in der Freizeit mit Whisky zu beschäftigen.

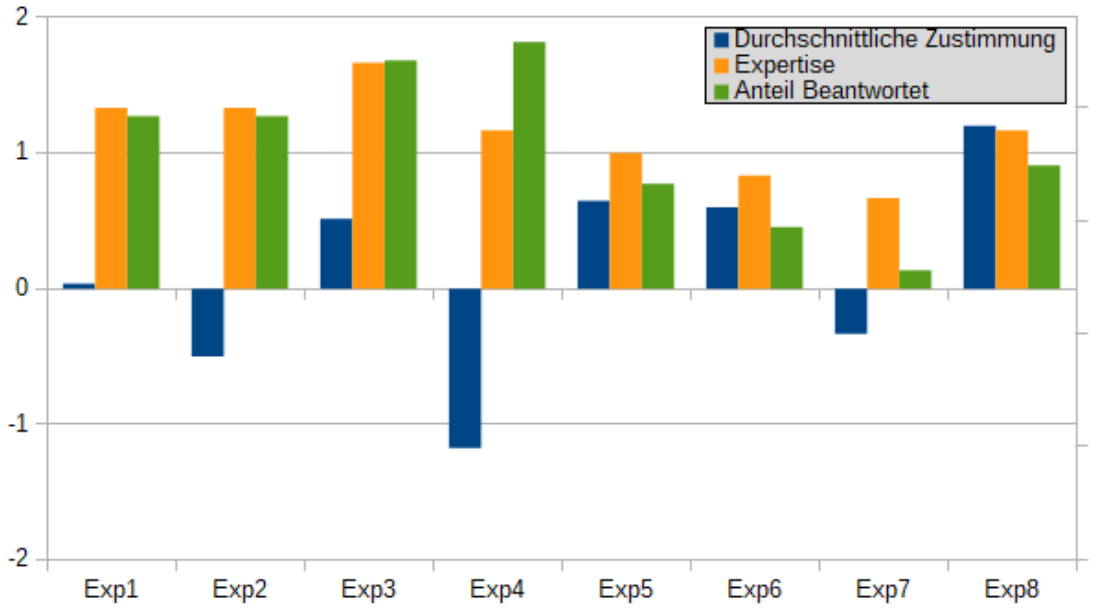


Abbildung 4.1.: Durchschnittliche Bewertungen des gesamten Fragebogens aufgeteilt nach Experten

Die Abbildung 4.1 zeigt zu jedem Experten die durchschnittliche Bewertung aller Fragen im Fragebogen. Dabei lässt sich erkennen, dass die Bewertungen zum Teil stark schwanken. Außerdem zeigt die Grafik den Anteil der beantworteten Fragen pro Experte an. Hier lässt sich klar erkennen, dass die vier Verkäufer einen deutlich höheren Anteil an beantworteten Fragen aufweisen. Dies bestätigt die Wichtigkeit echter Experten für diese Evaluation. Eine weitere

interessante Beobachtung ist, dass sich die Einschätzung der eigenen Expertise und der Anteil der beantworteten Fragen teilweise decken.

Die durchschnittliche Zustimmung über alle beantworteten Fragen hinweg beträgt $-0,01$ auf einer Skala von -2 bis $+2$. Dabei fällt die Zustimmung zu den Empfehlungen mit $0,03$ besser aus als die zu den präsentierten Auswahlen mit $-0,3$. Eine Gewichtung der Antworten anhand der Expertise der Teilnehmer führt zu einer leichten Verbesserung der gesamten Bewertung auf $0,01$. Die Bewertung der Empfehlungen steigt dabei auf $0,04$, die der Auswahlen auf $-0,2$. Insgesamt lässt sich die Bewertung durch die Experten als ausgeglichen bezeichnen. Hervorzuheben sind die teilweise deutlich schlechteren Bewertungen der Auswahlen. Ein Teil der schlechten Bewertungen hat ihre Ursache gegebenenfalls in der Gestaltung des Fragebogens. Es ist möglich, dass die Intention gerade hinter den präsentierten Auswahlen nicht deutlich genug beschrieben ist.

Zusätzliches Feedback

Im direkten Gespräch mit den Experten wurden bereits die ersten Probleme mit den Empfehlungen genannt. Eine Aussage ist, dass Whiskys aus verschiedenen Rohstoffen sehr unterschiedliche Aromen haben, was in den Empfehlungen aber nicht abgebildet würde. Eine weitere Aussage ist, dass der preisliche Unterschied zwischen den einzelnen Whiskys zu groß sei, was für ein Produktivsystem ungeeignet wäre. Weiterhin wurde gesagt, dass eine alleinige Betrachtung der Tasting Notes die Komplexität des Bereichs Whisky nicht ausreichend erfasse. Zudem wird bemängelt, dass einige Whiskys im Datenset auftauchen, die sich aktuell nicht mehr auf dem Markt befinden und mit Bezug auf die Cluster-Auswahlen, dass für ein Tasting in der Regel sechs Whiskys angeboten werden. Die Nichtverfügbarkeit einiger Whiskys stellt eine mögliche Erklärung für einen Teil der schlechten Bewertungen dar. So ist es denkbar, dass der Vorschlag eines nicht verfügbaren Whiskys zu der Antwort „nicht nachvollziehbar“führt.

Die fehlende Distinktion zwischen Whiskys aus unterschiedlichen Rohstoffen ist eventuell damit erklärbar, dass die Verfasser der Tasting Notes möglicherweise dazu neigen, das offensichtliche nicht mit in den Text aufzunehmen. Dazu würde unter anderem der verwendete Rohstoff gehören. Andererseits ist dies ein mögliches Indiz dafür, dass die Experten zu sehr in den bestehenden Kategorien denken. Die zu große Preisdifferenz ist für diese Arbeit kein entscheidender Faktor, da Metadaten dieser Art bewusst ignoriert wurden und das Ziel nicht darin besteht, Empfehlungen zu generieren, welche direkt in einem Produktivsystem verwendet werden könnten.

4.3. Möglichkeiten zur Verbesserung des Systems

Aufgrund der im vorigen Kapitel genannten Aussagen der Experten ergeben sich einige Möglichkeiten, die Empfehlungen zu verbessern. Am naheliegendsten erscheint die Maßnahme, den jeweiligen Rohstoff eines Whiskys mit einzubeziehen. Dies ist allerdings ebenfalls komplex, da das Verhältnis der verwendeten Getreidesorten teilweise variiert. Vor dieser Maßnahme wäre eine andere, erfolgversprechende Maßnahme die Berücksichtigung von quantifizierenden Wörtern. Dadurch würde das Verhältnis der Aromen zueinander abgebildet, was bisher nur durch Mehrfachnennungen geschieht. Nach dieser Maßnahme müsste eine erneute Befragung der Experten stattfinden, um zu sehen, ob sich der genannte Makel eventuell bereits gelöst hat. Das Risiko, dass in den Tasting Notes das offensichtliche dennoch ausgelassen wird, besteht dabei weiterhin. Deswegen ist es einigermaßen wahrscheinlich, dass das Einbeziehen der Rohstoffe dennoch nötig wäre.

Wie bereits erwähnt wäre eine Berücksichtigung von Alter und Herkunft nur in Kombination hilfreich. Hierbei gestaltet sich allerdings das Problem, dass gerade das Alter teilweise nicht angegeben wird.

Eine weitere, sehr vielversprechende Maßnahme wäre das Zusammenlegen aller Datensätze für Training und Testing. Auf diese Weise würde sich die Anzahl der Tasting Notes pro Whisky deutlich erhöhen. Ein Problem dabei ist jedoch, dass die Datensätze, die den gleichen Whisky beschreiben einander zugeordnet werden müssen, was sich als nicht trivial herausgestellt hat (Schole, 2017b). Eine Möglichkeit, diesen Prozess zu vereinfachen, wäre es, Whisky-Datensätze allein aufgrund ihrer Namen zusammenzufassen. Die somit erhöhte Anzahl an Tasting Notes pro Whisky würde außerdem die Notwendigkeit der Maßnahme, quantifizierende Begriffe zu betrachten, verringern, da mehrfach genannte Begriffe automatisch einen höheren Einfluss auf ein Whisky-Embedding haben. Die Erweiterung des Testdatensets um weitere Ressourcen hätte ebenfalls möglicherweise zur Folge, dass das Verhältnis von Scotch und anderen Whisky-Sorten verbessert würde. Momentan nehmen schottische Whiskys etwa 75% des Testdatensets ein. Dies kann zu einer Verzerrung bei den Empfehlungen und beim Clustering führen. Es ist allerdings nicht sicher, dass das Mengenverhältnis der Whisky-Sorten bei anderen Ressourcen gleichmäßiger ist. Schottischer Whisky ist in der Domäne generell dominant.

Eine Maßnahme, die die Anwendbarkeit der Empfehlungen in einem Produktivsystem erhöhen würde, wäre der Ausschluss von nicht mehr verfügbaren Whiskys aus dem Testdatenset. Zudem legen die erhaltenen Rückmeldungen nahe, dass im Falle eines weiteren Durchlaufs des Prozesses zu einem deutlich früheren Zeitpunkt mindestens ein Experte zu Rate gezogen

werden sollte, um Fehler wie die Miteinbeziehung von veralteten Whiskys frühzeitig zu eliminieren. Diese Möglichkeit ist allerdings insofern kritisch zu betrachten, dass Experten, die bereit sind, den gesamten Prozess zu betreuen, schwer zu finden sind. Zudem birgt die frühzeitige Einbeziehung die Gefahr, dass die Experten indirekt oder direkt einen zu großen Einfluss auf den Prozess an sich nehmen. So hätten die genannten Aussagen gegebenenfalls dazu geführt, dass die verwendeten Rohstoffe von Beginn an mit einbezogen worden wären und somit eine Durchführung ohne diese gar nicht stattgefunden hätte.

Eine offensichtliche Methode, die Qualität der Empfehlungen zu verbessern, ist die Beschaffung von noch mehr Datensätzen für das Training.

5. Fazit und Ausblick

Diese Arbeit befasst sich mit der Frage, ob allein auf Tasting Notes basierend nachvollziehbare Whisky-Empfehlungen gegeben werden können. Hierfür wird zunächst ein angepasster KDD-Prozess umrissen. Zu Beginn dieses Prozesses findet eine eingehende Betrachtung der Domäne Whisky mit ihren Eigenschaften statt. Dabei fällt bereits auf, dass dieser Gegenstandsbereich sehr komplex ist. Im Anschluss an die Betrachtung der Domäne findet eine Recherche möglicher Datenressourcen statt. Danach wird auf den Ergebnissen der Recherche basierend ein Datenkorpus aus Tasting Notes und den dazugehörigen Metadaten des jeweiligen beschriebenen Whiskys aufgebaut. Durch das Training eines Neuronalen Netzes mit dem Word2Vec-Algorithmus entstehen in der Folge Word Embeddings, welche die Ähnlichkeiten von Begriffen aus dem Textkorpus abbilden. Diese Word Embeddings bilden damit eine Möglichkeit, die Distanzen zwischen Geschmäckern und Aromen darzustellen. Es finden einige Versuche zur Optimierung des Word Embeddings mit gängigen Methoden aus dem Text-Preprocessing statt. Diese führen zu einer Verfeinerung der Word Embeddings. Im Anschluss folgt eine Anwendung der Embeddings auf das Testdatenset. Dies geschieht durch die Ermittlung des Mittelvektors der Embeddings eines Whiskys. Der dadurch entstandene Vektor dient als Repräsentationsform der Whiskys für die weitere Verarbeitung. Über die Distanzen der Whisky-Vektoren können Empfehlungen zu Whiskys genannt werden. Es folgt ein Clustering der Whiskys anhand ihrer Repräsentationsvektoren. Dabei wird die Tendenz beobachtet, dass zumindest besonders rauchige Whiskys einen der Cluster bilden. Die Qualität der Empfehlungen und Cluster muss unter Zuhilfenahme einiger Experten evaluiert werden. Die Ergebnisse dieser Befragung deuten auf eine mittlere Akzeptanz bei den Experten hin. Dabei schwanken die Einschätzungen der Teilnehmer jedoch stark.

Im Anschluss an die Evaluation findet eine Betrachtung möglicher Verbesserungen am Prozess statt. Zu diesen gehört unter anderem die Miteinbeziehung einiger Metadaten wie dem verwendeten Rohstoff eines Whiskys in die Empfehlungsgenerierung. Dies stünde allerdings im Gegensatz zur ursprünglichen Fragestellung dieser Arbeit. Des weiteren bietet sich die Betrachtung quantifizierender Begriffe an, um eine realistischere Gewichtung der Geschmacksnoten zu

erreichen. Eine Vergrößerung des Datenkorpus stellt immer eine Möglichkeit zur Verbesserung des Systems dar.

Insgesamt steht die Erkenntnis, dass die in dieser Arbeit generierten Whisky-Empfehlungen mittelmäßig akzeptabel sind.

A. Texte nach diversen Vorverarbeitungsschritten

A.1. Beispieltexte in Rohform

Quelle	Beispieltext
Distiller (2018)	The aroma is full of tropical golden fruits with a fairly decent intensity. Toasted coconut and sweet curry spices join in the fun. The palate shows much of the same with a nice balance between malt and fruit with the sherry casks heightening and not overwhelming the whisky. A trace of char smoke and salt is present, but only just. The finish is lovely and lasting.
Scotchwhisky.com (2018)	Soft, almost oily, mature mix of old column still rum, some char alongside white chocolate, ripe banana/banana chews and red fruit. Greener with water.. Sweet start with those enthusiastic rummy elements, and a firm back-palate. Light chocolate (darker now) alongside raspberry. Soft, sweet and mature, but there's power here. Water reduces the complexity a little.. Slightly sharp.
Whisky Advocate (2018)	Angela D'Orazio matured a third of the component whiskies in cherry-wine casks for this seasonal creation. Rather than cherries, the fruity aroma of the wine is more apparent on the nose, together with crushed root ginger, hawthorn, drying tobacco leaf, and beefsteak tomatoes. This thick-textured dram exudes cherry, strawberry, and vanilla fudge, drawing the mouth before a late phase of rum and raisin, aniseed, peppermint, and menthol. 598 SEK

Whisky Magazine (2018)	Notably sweet, but also tea-like.. Satiny. Starts drier. More tightly combined flavours take longer to unfold. Gradually sweeter and more orangey (or tropical fruit?), but always with burnt-grass peatiness behind. Much more complex.. Very long, soothing, warming.
Whiskyology (2018)	Fragrant wood, rich pot still with a touch of sherry sweetness.. All the creaminess and honey sweetness come together to create a taste sensation of spicy character, a smooth complex whiskey experience.. A very long finish with a pleasantly peppery finale.
Brossard (2018)	Nose: Juicy, peaty, maritime, mineral, slightly fruity, with some iodine, seaweed, blackcurrant and disinfectant.. Taste: Fruity, smoky, slightly peaty and mineral, on intense blackcurrant juice, some iodine, eucalyptus, seaweeds, tar, ashes and peat smoke, as well as faint notes of soot. The finish is medium, juicy, oaky, fruity, rather sweet, slightly mineral and smoky, on blackcurrant, some passion fruits, iodine, strawberries, seaweeds, tar and ashes..
Klaverstijn (2018)	Nose: Lovely stuff! Very much in the same ballpark as all those 21-year-olds I tasted late last year. The sherry influence is subtle and very well-integrated. Plum juice, raspberries and sweet oranges. It has a lovely minerality to it, as well as a slight sootiness, a whisper of polished leather and soft notes of cigar tobacco. Finally some brown sugar, icing and after a little while even a hint of soy. Top notch... Taste: Oh yes! Is this really almost 60 percent? Lovely, juicy fruits. Apricot jam, strawberries, and oranges. Somewhat waxy, a touch of brine and gentle notes of peat. Also some furniture polish, as well as menthol and a touch of licorice. Very classy. Finish: Basically all of the above, with the fruits lingering. Medium in length.

Malt (2018)	Colour: pine wood. On the nose: green apples, vanilla and sour cream initially, it's all very fresh and lively. Icing sugar and cotton sheets, almonds and some lemon sponge cake. Coconut towards the rear rounds off a classic Speyside set of aromas. In the mouth: it's all very crisp with those apples once again and digestive biscuits. Lemon peelings, a lime cordial with more of the coconut and the drying bitterness from the wood..
Thomson (2018)	Utter bliss on the nose, dewy grass, light watery fruits like melon and nectarines slight vanilla and touch of light wood. coconut milk, slightly chalky, vanilla pods, nose of the grassy notes i usually find on the palette of grains. coconut and vanilla custard
Lardin (2018)	strong. floral, honey, orange zest, spices, peaches, wood. smooth, oily, powerful. floral, spices, honey, wood, liquorice. long. honey, floral, citrus, spices.

Whisky Intelligence (2018)	Caol Ila 12yo 1992/2005 (50%, Douglas Laing OMC, C#1830, 388 Bts., D10/'92 B05/'05) Another single cask bottling by Douglas Laing & Co, without chill filtration and 'no colouring' and from a refill hogshead. The nose reveals good peat smoke, burnt twigs, hints of Dettol and iodine. So far it's an Islay lovers delight. There is also cocoa, wafts of a lumber yard (think of piles of saw dust and freshly cut cedar), malt dust, sweet vanilla and some hints of junipers. Nice and solid. The taste is a crescendo of peat smoke backed by the juniper and then the sweetness along with some apples, the burnt twigs an then some more peat smoke. Excellent! Some water creates whirls and eddies in the glass and changes the whisky some; smoother and sweeter but the peat smoke is still the dominate characteristic and it's very good. A little water goes a long way and improves it quite a bit. The finish is creamy at first, peated and then wood spice along with Fry's unsweetened cocoa and then the creaminess once again. It is big, long and warming. There is loads of peat smoke for those that want it. After nearly 10 minutes the finish is still quite evident but has now changed to elastoplast bandages. A cracker of a Caol Ila and shows the brilliance of a refill hogs head and peat smoke. Score 88 points Many thanks for the sample Rich!
Whisky Monitor (2018)	Baked oranges, sprinkled with cinnamon. Red ripe apples. Raisin appears, quite a bit. Tasty, quite in balance with the nose. Raisin, baked ripe fruits, sweet. Quite powerful aftertaste. Handles water well.

Whisky Notes (2018)	Nose: very bold bourbon oak influence. Different sorts of warm wood, sandalwood, cedar from cigar boxes, some thuja... In fact I like this kind of oakiness, it's elegant and matches the oriental character. There's also varnish and solventy notes. Leather. Underneath it has apricot jam, yellow plums and vanilla-coated berry cake. Touches of mint, with floral overtones. Mouth: this is where the wood starts to show more astrigency. Fruits are now heavily infused fruit tea. Slightly tangy ginger and clove as well. A little coconut oil. Unfortunately also a planky note which coats your mouth and a little tobacco sourness. Orange peel. Spruce needles. Again some flowery touches. Finish: long, quite floral and heady. Mint and traces of the apricot jam.

Whiskey Reviewer (2018)	Label 5 Gold Heritage has a light copper color in the glass, with small tears. On the nose is fresh and grassy. There are fresh fruit notes, a classic Label 5, among them you can smell some baked apples and a sassy touch of oranges. There are also light notes of malted cereals, that come along with sweet vanilla and honey aromas. It has also a gentle touch of cinnamon., The whisky is mellow and light-bodied. There is some smoothness at first on the palate but later becomes a more complex dram. Label 5 Gold Heritage starts as a sweet and fruity blend. There are some caramel and vanilla notes in the beginning, followed by apples and oranges. Then, there is a grainy touch followed by an explosion of spices. There are cinnamon and ginger, with some faint touches of smoke., The end is short and dry. There are some lingering spicy notes mixed up with a little touch of oak., Label 5 Gold Heritage keeps the Label 5 character, with fresh fruity notes and a mellow and well-balanced dram. Even though, the aromas are more promising that what we finally find on the palate., This expression doesn't keep up with the great quality-price ratio that we can find in other blends from Label 5's range. Although it is not as interesting as Label 5 Classic Black or its 12 Year Old brother, it is a nice and fresh dram, perfect for cocktails and warmer weather.

A.2. Beispieltexte nach Import und Konversion in Kleinbuchstaben

Quelle	Beispieltext
Distiller (2018)	the aroma is full of tropical golden fruits with a fairly decent intensity. toasted coconut and sweet curry spices join in the fun. the palate shows much of the same with a nice balance between malt and fruit with the sherry casks heightening and not overwhelming the whisky. a trace of char smoke and salt is present but only just. the finish is lovely and lasting.

Scotchwhisky.com (2018)	soft almost oily mature mix of old column still rum some char alongside white chocolate ripe banana banana chews and red fruit. greener with water sweet start with those enthusiastic rummy elements and a firm back palate. light chocolate darker now alongside raspberry. soft sweet and mature but there's power here. water reduces the complexity a little slightly sharp.
Whisky Advocate (2018)	angela d'orazio matured a third of the component whiskies in cherry wine casks for this seasonal creation. rather than cherries the fruity aroma of the wine is more apparent on the nose together with crushed root ginger hawthorn drying tobacco leaf and beefsteak tomatoes. this thick textured dram exudes cherry strawberry and vanilla fudge drawing the mouth before a late phase of rum and raisin aniseed peppermint and menthol. 598 sek.
Whisky Magazine (2018)	notably sweet but also tea like satiny. starts drier. more tightly combined flavours take longer to unfold. gradually sweeter and more orangey or tropical fruit. but always with burnt grass peatiness behind. much more complex. very long soothing warming.
Whiskyology (2018)	fragrant wood rich pot still with a touch of sherry sweetness all the creaminess and honey sweetness come together to create a taste sensation of spicy character a smooth complex whiskey experience a very long finish with a pleasantly peppery finale.
Brossard (2018)	nose juicy peaty maritime mineral slightly fruity with some iodine seaweed blackcurrant and disinfectant. taste fruity smoky slightly peaty and mineral on intense blackcurrant juice some iodine eucalyptus seaweeds tar ashes and peat smoke as well as faint notes of soot. the finish is medium juicy oaky fruity rather sweet slightly mineral and smoky on blackcurrant some passion fruits iodine strawberries seaweeds tar and ashes.

Klaverstijn (2018)	nose lovely stuff. very much in the same ballpark as all those 21 year olds i tasted late last year. the sherry influence is subtle and very well integrated. plum juice raspberries and sweet oranges. it has a lovely minerality to it as well as a slight sootiness a whisper of polished leather and soft notes of cigar tobacco. finally some brown sugar icing and after a little while even a hint of soy. top notch. taste oh yes. is this really almost 60 percent. lovely juicy fruits. apricot jam strawberries and oranges. somewhat waxy a touch of brine and gentle notes of peat. also some furniture polish as well as menthol and a touch of licorice. very classy. finish basically all of the above with the fruits lingering. medium in length.
Malt (2018)	colour pine wood. on the nose green apples vanilla and sour cream initially it's all very fresh and lively. icing sugar and cotton sheets almonds and some lemon sponge cake. coconut towards the rear rounds off a classic speyside set of aromas. in the mouth it's all very crisp with those apples once again and digestive biscuits. lemon peelings a lime cordial with more of the coconut and the drying bitterness from the wood.
Thomson (2018)	utter bliss on the nose dewy grass light watery fruits like melon and nectarines slight vanilla and touch of light wood. coconut milk slightly chalky vanilla pods nose of the grassy notes i usually find on the palette of grains. coconut and vanilla custard.
Lardin (2018)	strong. floral honey orange zest spices peaches wood. smooth oily powerful. floral spices honey wood liquorice. long. honey floral citrus spices.

Whisky Intelligence (2018)	caol ila 12yo 1992 2005 50 douglas laing omc c 1830 388 bts d10 '92 b05 '05 another single cask bottling by douglas laing co without chill filtration and 'no colouring' and from a refill hogshead. the nose reveals good peat smoke burnt twigs hints of dettol and iodine. so far it's an islay lovers delight. there is also cocoa wafts of a lumber yard think of piles of saw dust and freshly cut cedar malt dust sweet vanilla and some hints of junipers. nice and solid. the taste is a crescendo of peat smoke backed by the juniper and then the sweetness along with some apples the burnt twigs an then some more peat smoke. excellent. some water creates whirls and eddies in the glass and changes the whisky some smoother and sweeter but the peat smoke is still the dominate characteristic and it's very good. a little water goes a long way and improves it quite a bit. the finish is creamy at first peated and then wood spice along with fry's unsweetened cocoa and then the creaminess once again. it is big long and warming. there is loads of peat smoke for those that want it. after nearly 10 minutes the finish is still quite evident but has now changed to elastoplast bandages. a cracker of a caol ila and shows the brilliance of a refill hogs head and peat smoke. score 88 points many thanks for the sample rich.
Whisky Monitor (2018)	baked oranges sprinkled with cinnamon. red ripe apples. raisin appears quite a bit. tasty quite in balance with the nose. raisin baked ripe fruits sweet. quite powerful aftertaste. handles water well.

Whisky Notes (2018)	nose very bold bourbon oak influence. different sorts of warm wood sandalwood cedar from cigar boxes some thuja. in fact i like this kind of oakiness it's elegant and matches the oriental character. there's also varnish and solventy notes. leather. underneath it has apricot jam yellow plums and vanilla coated berry cake. touches of mint with floral overtones. mouth this is where the wood starts to show more astrigency. fruits are now heavily infused fruit tea. slightly tangy ginger and clove as well. a little coconut oil. unfortunately also a planky note which coats your mouth and a little tobacco sourness. orange peel. spruce needles. again some flowery touches. finish long quite floral and heady. mint and traces of the apricot jam.
Whiskey Reviewer (2018)	label 5 gold heritage has a light copper color in the glass with small tears. on the nose is fresh and grassy. there are fresh fruit notes a classic label 5 among them you can smell some baked apples and a sassy touch of oranges. there are also light notes of malted cereals that come along with sweet vanilla and honey aromas. it has also a gentle touch of cinnamon the whisky is mellow and light bodied. there is some smoothness at first on the palate but later becomes a more complex dram. label 5 gold heritage starts as a sweet and fruity blend. there are some caramel and vanilla notes in the beginning followed by apples and oranges. then there is a grainy touch followed by an explosion of spices. there are cinnamon and ginger with some faint touches of smoke the end is short and dry. there are some lingering spicy notes mixed up with a little touch of oak label 5 gold heritage keeps the label 5 character with fresh fruity notes and a mellow and well balanced dram. even though the aromas are more promising that what we finally find on the palate this expression doesn't keep up with the great quality price ratio that we can find in other blends from label 5's range. although it is not as interesting as label 5 classic black or its 12 year old brother it is a nice and fresh dram perfect for cocktails and warmer weather.

A.3. Beispieltexte nach Stopword Removal

Quelle	Beispieltext
Distiller (2018)	aroma full tropical golden fruits fairly decent intensity. toasted coconut sweet curry spices join fun. palate shows much nice balance malt fruit sherry casks heightening overwhelming whisky. trace char smoke salt present. finish lovely lasting.
Scotchwhisky.com (2018)	soft almost oily mature mix old column still rum char alongside white chocolate ripe banana banana chews red fruit. greener water sweet start enthusiastic rummy elements firm back palate. light chocolate darker alongside raspberry. soft sweet mature there's power. water reduces complexity little slightly sharp.
Whisky Advocate (2018)	angela d'orazio matured third component whiskies cherry wine casks seasonal creation. rather cherries fruity aroma wine apparent nose together crushed root ginger hawthorn drying tobacco leaf beefsteak tomatoes. thick textured dram exudes cherry strawberry vanilla fudge drawing mouth late phase rum raisin aniseed peppermint menthol. 598 sek.
Whisky Magazine (2018)	notably sweet also tea like satiny. starts drier. tightly combined flavours take longer unfold. gradually sweeter orangey tropical fruit. always burnt grass peatiness behind. much complex. long soothing warming.
Whiskyology (2018)	fragrant wood rich pot still touch sherry sweetness creaminess honey sweetness come together create taste sensation spicy character smooth complex whiskey experience long finish pleasantly peppery finale.
Brossard (2018)	nose juicy peaty maritime mineral slightly fruity iodine seaweed blackcurrant disinfectant. taste fruity smoky slightly peaty mineral intense blackcurrant juice iodine eucalyptus seaweeds tar ashes peat smoke well faint notes soot. finish medium juicy oaky fruity rather sweet slightly mineral smoky blackcurrant passion fruits iodine strawberries seaweeds tar ashes.

Klaverstijn (2018)	nose lovely stuff. much ballpark 21 year olds tasted late last year. sherry influence subtle well integrated. plum juice raspberries sweet oranges. lovely minerality well slight sootiness whisper polished leather soft notes cigar tobacco. finally brown sugar icing little even hint soy. top notch. taste oh yes. really almost 60 percent. lovely juicy fruits. apricot jam strawberries oranges. somewhat waxy touch brine gentle notes peat. also furniture polish well menthol touch licorice. classy. finish basically fruits lingering. medium length.
Malt (2018)	colour pine wood. nose green apples vanilla sour cream initially fresh lively. icing sugar cotton sheets almonds lemon sponge cake. coconut towards rear rounds classic speyside set aromas. mouth crisp apples digestive biscuits. lemon peelings lime cordial coconut drying bitterness wood.
Thomson (2018)	utter bliss nose dewy grass light watery fruits like melon nectarines slight vanilla touch light wood. coconut milk slightly chalky vanilla pods nose grassy notes usually find palette grains. coconut vanilla custard.
Lardin (2018)	strong. floral honey orange zest spices peaches wood. smooth oily powerful. floral spices honey wood liquorice. long. honey floral citrus spices.

Whisky Intelligence (2018)	caol ila 12yo 1992 2005 50 douglas laing omc c 1830 388 bts d10 ’92 b05 ’05 another single cask bottling douglas laing co without chill filtration ’no colouring’ refill hogshead. nose reveals good peat smoke burnt twigs hints dettol iodine. far islay lovers delight. also cocoa wafts lumber yard think piles saw dust freshly cut cedar malt dust sweet vanilla hints junipers. nice solid. taste crescendo peat smoke backed juniper sweetness along apples burnt twigs peat smoke. excellent. water creates whirls eddies glass changes whisky smoother sweeter peat smoke still dominate characteristic good. little water goes long way improves quite bit. finish creamy first peated wood spice along fry’s unsweetened cocoa creaminess. big long warming. loads peat smoke want. nearly 10 minutes finish still quite evident changed elastoplast bandages. cracker caol ila shows brilliance refill hogs head peat smoke. score 88 points many thanks sample rich.
Whisky Monitor (2018)	baked oranges sprinkled cinnamon. red ripe apples. raisin appears quite bit. tasty quite balance nose. raisin baked ripe fruits sweet. quite powerful aftertaste. handles water well.
Whisky Notes (2018)	nose bold bourbon oak influence. different sorts warm wood sandalwood cedar cigar boxes thuja. fact like kind oakiness elegant matches oriental character. there’s also varnish solventy notes. leather. underneath apricot jam yellow plums vanilla coated berry cake. touches mint floral overtones. mouth wood starts show astringency. fruits heavily infused fruit tea. slightly tangy ginger clove well. little coconut oil. unfortunately also planky note coats mouth little tobacco sourness. orange peel. spruce needles. flowery touches. finish long quite floral heady. mint traces apricot jam.

Whiskey Reviewer (2018)	label 5 gold heritage light copper color glass small tears. nose fresh grassy. fresh fruit notes classic label 5 among smell baked apples sassy touch oranges. also light notes malted cereals come along sweet vanilla honey aromas. also gentle touch cinnamon whisky mellow light bodied. smoothness first palate later becomes complex dram. label 5 gold heritage starts sweet fruity blend. caramel vanilla notes beginning followed apples oranges. grainy touch followed explosion spices. cinnamon ginger faint touches smoke end short dry. lingering spicy notes mixed little touch oak label 5 gold heritage keeps label 5 character fresh fruity notes mellow well balanced dram. even though aromas promising finally find palate expression keep great quality price ratio find blends label 5's range. although interesting label 5 classic black 12 year old brother nice fresh dram perfect cocktails warmer weather.

A.4. Beispieltexte nach Stemming

Quelle	Beispieltext
Distiller (2018)	aroma full tropic golden fruit fair decent inten. toast coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri cask heighten overwhelm whiski. trace char smoke salt present. finish love last.
Scotchwhisky.com (2018)	soft almost oili matur mix old column still rum char alongsid white chocol ripe banana banana chew red fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.
Whisky Advocate (2018)	angela d'orazio matur third compon whiski cherri wine cask season creation. rather cherri fruiti aroma wine appar nose togeth crush root ginger hawthorn dri tobacco leaf beefsteak tomato. thick textur dram exud cherri strawberri vanilla fudg draw mouth late phase rum raisin anis peppermint menthol. 598 sek.

Whisky Magazine (2018)	notabl sweet also tea like satini. start drier. tight combin flavour take longer unfold. gradual sweeter orangey tropic fruit. alway burnt grass peati behind. much complex. long sooth warm.
Whiskyology (2018)	fragrant wood rich pot still touch sherri sweet creami honey sweet come togeth creat tast sensat spici charact smooth complex whiskey experi long finish pleasant pepperi final.
Brossard (2018)	nose juici peati maritim miner slight fruiti iodin seaw blackcurr disinfect. tast fruiti smoki slight peati miner inten blackcurr juic iodin eucalyptus seaw tar ash peat smoke well faint note soot. finish medium juici oaki fruiti rather sweet slight miner smoki blackcurr passion fruit iodin strawberri seaw tar ash.
Klaverstijn (2018)	nose love stuff. much ballpark 21 year old tast late last year. sherri influenc subtl well integr. plum juic raspberri sweet orang. love miner well slight sooti whisper polish leather soft note cigar tobacco. final brown sugar ice littl even hint soy. top notch. tast oh yes. realli almost 60 percent. love juici fruit. apricot jam strawberri orang. somewhat waxi touch brine gentl note peat. also furnitur polish well menthol touch licor. classi. finish basic fruit linger. medium length.
Malt (2018)	colour pine wood. nose green appl vanilla sour cream initi fresh live. ice sugar cotton sheet almond lemon spong cake. coconut toward rear round classic speysid set aroma. mouth crisp appl digest biscuit. lemon peel lime cordial coconut dri bitter wood.
Thomson (2018)	utter bliss nose dewi grass light wateri fruit like melon nectarin slight vanilla touch light wood. coconut milk slight chalki vanilla pod nose grassi note usual find palett grain. coconut vanilla custard.
Lardin (2018)	strong. floral honey orang zest spice peach wood. smooth oili power. floral spice honey wood liquor. long. honey floral citrus spice.

Whisky Intelligence (2018)	caol ila 12yo 1992 2005 50 dougla la omc c 1830 388 bts d10 92 b05 05 anoth singl cask bottl dougla la co without chill filtrat no colour refil hogshead. nose reveal good peat smoke burnt twig hint dettol iodin. far islay lover delight. also cocoa waft lumber yard think pile saw dust fresh cut cedar malt dust sweet vanilla hint junip. nice solid. tast crescendo peat smoke back junip sweet along appl burnt twig peat smoke. excel. water creat whirl eddi glass chang whiski smoother sweeter peat smoke still domin characterist good. littl water goe long way improv quit bit. finish creami first peat wood spice along fri unsweeten cocoa creami. big long warm. load peat smoke want. near 10 minut finish still quit evid chang elastoplast bandag. cracker caol ila show brillianc refil hog head peat smoke. score 88 point mani thank sampl rich.
Whisky Monitor (2018)	bake orang sprinkl cinnamon. red ripe appl. raisin appear quit bit. tasti quit balanc nose. raisin bake ripe fruit sweet. quit power aftertast. handl water well.
Whisky Notes (2018)	nose bold bourbon oak influenc. differ sort warm wood sandalwood cedar cigar box thuja. fact like kind oaki eleg match orient charact. there also varnish solventi note. leather. underneath apricot jam yellow plum vanilla coat berri cake. touch mint floral overton. mouth wood start show astrig. fruit heavili infus fruit tea. slight tangi ginger clove well. littl coconut oil. unfortun also planki note coat mouth littl tobacco sour. orang peel. spruce needl. floweri touch. finish long quit floral headi. mint trace apricot jam.

Whiskey Reviewer (2018)	label 5 gold heritag light copper color glass small tear. nose fresh grassi. fresh fruit note classic label 5 among smell bake appl sassi touch orang. also light note malt cereal come along sweet vanilla honey aroma. also gentl touch cinnamon whiski mellow light bodi. smooth first palat later becom complex dram. label 5 gold heritag start sweet fruiti blend. caramel vanilla note begin follow appl orang. graini touch follow explo spice. cinnamon ginger faint touch smoke end short dri. linger spici note mix littl touch oak label 5 gold heritag keep label 5 charact fresh fruiti note mellow well balanc dram. even though aroma promi final find palat express keep great qualiti price ratio find blend label 5 rang. although interest label 5 classic black 12 year old brother nice fresh dram perfect cocktail warmer weather.

A.5. Beispieltexte nach Phrase Detection

Quelle	Beispieltext
Distiller (2018)	aroma full tropic golden fruit fair decent inten. toast_coconut sweet curri spice join fun. palat show much nice balanc malt fruit sherri_cask heighten overwhelm whiski. trace char smoke salt present. finish love last.
Scotchwhisky.com (2018)	soft almost oili matur mix old column_still rum char alongsid white_chocol ripe_banana banana chew red_fruit. greener water sweet start enthusiast rummi element firm back palat. light chocol darker alongsid raspberri. soft sweet matur there power. water reduc complex littl slight sharp.
Whisky Advocate (2018)	angela d'orazio matur third compon whiski cherri wine_cask season creation. rather cherri fruiti aroma wine appar nose togeth crush root_ginger hawthorn dri tobacco_leaf beefsteak tomato. thick_textur dram exud cherri strawberri vanilla fudg draw mouth late phase rum_raisin anis peppermint menthol. 598 sek.

Whisky Magazine (2018)	notabl sweet also tea like satini. start drier. tight_combin_flavour take longer unfold. gradual sweeter orangey tropic_fruit. alway burnt_grass peati behind. much complex. long sooth warm.
Whiskyology (2018)	fragrant wood rich pot_still touch sherri sweet creami honey sweet come_togeth creat tast sensat spici charact smooth complex whiskey experi long finish pleasant pepperi final.
Brossard (2018)	nose juici peati_maritim miner slight fruiti iodin_seaw blackcurr disinfect. tast fruiti smoki slight peati miner inten blackcurr juic iodin eucalyptus seaw_tar_ash peat_smoke well faint note soot. finish_medium juici oaki fruiti rather sweet slight miner smoki blackcurr passion_fruit iodin strawberri seaw_tar_ash.
Klaverstijn (2018)	nose love_stuff. much ballpark 21_year_old tast late last_year. sherri_influenc subtl well_integr. plum juic raspberri sweet orang. love miner well slight sooti whisper polish_leather soft note cigar_tobacco. final brown_sugar ice littl even hint soy. top_notch. tast oh_yes. realli almost 60_percent. love juici fruit. apricot_jam strawberri orang. somewhat waxi touch brine gentl note peat. also furnitur_polish well menthol touch licor. classi. finish basic fruit linger. medium_length.
Malt (2018)	colour pine wood. nose green_appl vanilla sour cream initi fresh live. ice_sugar cotton_sheet almond lemon spong_cake. coconut toward rear round classic_speysid set aroma. mouth crisp appl digest_biscuit. lemon_peel lime_cordial coconut dri bitter wood.
Thomson (2018)	utter bliss nose dewi_grass light wateri fruit like melon nectarin slight vanilla touch light wood. coconut_milk slight chalki vanilla_pod nose grassi note usual find palett grain. coconut vanilla_custard.
Lardin (2018)	strong. floral honey orang_zest spice peach wood. smooth oili_power. floral spice honey wood liquor. long. honey floral citrus spice.

Whisky Intelligence (2018)	caol_ila 12yo 1992 2005 50 dougla_la omc c 1830 388 bts d10 92 b05 05 anoth singl_cask_bottl dougla_la_co without_chill_filtrat no colour refil_hogshead. nose reveal good peat_smoke burnt_twig hint dettol iodin. far islay lover delight. also cocoa waft lumber_yard think pile saw_dust fresh_cut cedar malt dust sweet vanilla hint junip. nice solid. tast crescendo peat_smoke back junip sweet along appl burnt_twig peat_smoke. excel. water creat whirl eddi glass chang whiski smoother sweeter peat_smoke still domin characterist good. littl water goe long way improv quit bit. finish creami first peat wood spice along fri unsweeten_cocoa creami. big long warm. load peat_smoke want. near 10_minut finish still quit evid chang elastoplast bandag. cracker caol_ila show brillianc refil hog_head peat_smoke. score_88_point mani_thank sampl rich.
Whisky Monitor (2018)	bake orang sprinkl cinnamon. red ripe appl. raisin appear quit bit. tasti quit balanc nose. raisin bake ripe fruit sweet. quit power aftertast. handl water well.
Whisky Notes (2018)	nose bold bourbon oak influenc. differ sort warm wood sandalwood cedar_cigar_box thuja. fact like kind oaki eleg match orient charact. there also varnish solventi note. leather. underneath apricot_jam yellow_plum vanilla coat berri cake. touch mint floral overton. mouth wood start show astrig. fruit heavili infus fruit tea. slight tangi ginger clove well. littl coconut_oil. unfortun also planki note coat_mouth littl tobacco sour. orang_peel. spruce needl. floweri touch. finish_long quit floral headi. mint trace apricot_jam.

Whiskey Reviewer (2018)	label_5 gold heritag light copper_color_glass small tear. nose fresh grassi. fresh fruit note classic label_5 among smell bake_appl sassi touch orang. also light note malt cereal come along sweet vanilla honey aroma. also gentl touch cinnamon whiski mellow light bodi. smooth first palat later becom complex dram. label_5 gold heritag start sweet fruiti blend. caramel vanilla note begin follow appl orang. graini touch follow explo spice. cinnamon ginger faint touch smoke end short dri. linger spici note mix littl touch oak label_5 gold heritag keep label_5 charact fresh fruiti note mellow well_balanc dram. even_though aroma promi final find palat express keep great qualiti price ratio find blend label_5 rang. although interest label_5 classic black 12_year_old brother nice fresh dram perfect cocktail warmer weather.

B. Fragebogen

B.1. Deutsche Variante

Umfrage zum Thema Whisky-Empfehlungen und -Klassifizierungen

Die Frage nach einer Whisky-Empfehlung stellt für den Befragten immer eine Herausforderung dar. Whisky ist ein großes Themenfeld mit einer Vielzahl an Ausprägungen. Daher wäre ein Software-Tool sehr hilfreich, das zuverlässig Empfehlungen für Whiskys aussprechen kann.

Ich bin Informatik-Student an der HAW Hamburg. Im Rahmen meiner Masterarbeit generiere ich Whisky-Empfehlungen durch den Vergleich von etwa 20.000 Tasting Notes mit Machine-Learning-Methoden.

Um die Ergebnisse meiner Arbeit zu evaluieren, benötige ich die Einschätzungen einiger Kenner und Experten, wie nachvollziehbar meine generierten Empfehlungen sind. Es geht dabei nicht darum, ob Sie diese Empfehlungen persönlich in gleicher Weise geben würden, sondern nur darum, ob Sie diese Empfehlungen als nachvollziehbar beschreiben würden.

Die Umfrage beansprucht in etwa 15 Minuten.

Kenntnisstand

Um die Umfrageergebnisse nach dem Kenntnisstand der Teilnehmer einordnen zu können, benötige ich zunächst eine Einschätzung Ihrer persönlichen Erfahrung.

Wie würden Sie Ihren persönlichen Kenntnisstand im Bereich Whisky auf einer Skala von 1 bis 7 - wobei 1 für Neuling und 7 für Experte steht - einschätzen?

	1	2	3	4	5	6	7	
Neuling	◯	◯	◯	◯	◯	◯	◯	Experte

Seit wie vielen Jahren setzen Sie sich ungefähr mit Whisky auseinander?

0-5	5-10	10-15	15-20	über 20
◯	◯	◯	◯	◯

In welcher Form beschäftigen Sie sich mit Whisky?

- ◯ in der Freizeit
- ◯ als Verkäufer
- ◯ als Kritiker

Ihre Angaben werden von mir selbstverständlich anonym behandelt und verarbeitet. Falls Sie dennoch wünschen, in der Danksagung meiner Thesis namentlich erwähnt zu werden, können Sie hier Ihren Namen angeben.

Empfehlungen

Im folgenden sind jeweils ein Referenz-Whisky und die dazu erzeugten Empfehlungen gezeigt. Der Referenz-Whisky steht hierbei beispielsweise für einen vom Kunden erfragten Whisky, zu dem möglichst ähnliche Empfehlungen genannt werden sollen. Bitte kreuzen Sie zu jeder Empfehlung an, wie sehr Sie diese als nachvollziehbar bezeichnen würden. Ich habe versucht, möglichst die bekanntesten Whiskys aus den wesentlichen bestehenden Kategorien als Referenz-Whiskys zu verwenden. Es handelt sich jeweils um die Fünf ähnlichsten Whiskys zum Referenz-Whisky. Sollten Sie den Referenz-Whisky nicht kennen, geben Sie dies bitte entsprechend an.

Die Bewertungsskala reicht hierbei von -2 (überhaupt nicht nachvollziehbar) bis 2 (sehr nachvollziehbar).

Referenz-Whisky: Lagavulin 16 Years Old

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Laphroaig 10 Years Old, Cask Strength	◯	◯	◯	◯	◯	◯
Lagavulin 12 Years Old	◯	◯	◯	◯	◯	◯
Laphroaig 10 Years Old	◯	◯	◯	◯	◯	◯
Ardbeg 10 Years Old	◯	◯	◯	◯	◯	◯
Laphroaig Quarter Cask	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Glenmorangie 10 Years Old

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Yamazaki 12 Years Old	◯	◯	◯	◯	◯	◯
The Glenlivet 12 Years Old	◯	◯	◯	◯	◯	◯
Black Mountain BM No2	◯	◯	◯	◯	◯	◯
Dewar's Signature	◯	◯	◯	◯	◯	◯
Midleton Single Cask 1996	◯	◯	◯	◯	◯	◯

Referenz-Whisky: The Balvenie 12 Years Old, Double Wood

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Speyburn 21 Years Old	◯	◯	◯	◯	◯	◯
Balblair 2004 Sherry Matured	◯	◯	◯	◯	◯	◯
Old Pulteney 1990 Vintage	◯	◯	◯	◯	◯	◯
Glen Moray 16 Years Old, Chenin Blanc Finish	◯	◯	◯	◯	◯	◯
Aberlour 1976	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Jameson Irish Whiskey

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Redbreast 12 Years Old	◯	◯	◯	◯	◯	◯
Glen Moray 16 Years Old, Chenin Blanc Finish	◯	◯	◯	◯	◯	◯
Speyburn 21 Years Old	◯	◯	◯	◯	◯	◯
Tullamore Dew 12 Years Old	◯	◯	◯	◯	◯	◯
The Wee Dram Balblair 12 Years Old	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Elijah Craig 12 Years Old

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Eagle Rare 17 Years Old	◯	◯	◯	◯	◯	◯
Talisker 30 Years Old	◯	◯	◯	◯	◯	◯
George T. Stagg	◯	◯	◯	◯	◯	◯
Old Forester Birthday Bourbon	◯	◯	◯	◯	◯	◯
Yamazaki 18 Years Old	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Talisker 10 Years Old

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
Talisker 25 Years Old	◯	◯	◯	◯	◯	◯
Brora 30 Years Old	◯	◯	◯	◯	◯	◯
Lagavulin 12 Years Old	◯	◯	◯	◯	◯	◯
Caol Ila Cask Strength	◯	◯	◯	◯	◯	◯
Black Bottle 10 Years Old	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Nikka From the Barrel

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
The Dalmore Cigar Malt	◯	◯	◯	◯	◯	◯
Tamdhu Batch Strength 002	◯	◯	◯	◯	◯	◯
Celtique Connexion Vin de Paille du Jura Wood Finish 1994	◯	◯	◯	◯	◯	◯
Girvan Single Grain 1964	◯	◯	◯	◯	◯	◯
Tomatin 14 Years Old 2002	◯	◯	◯	◯	◯	◯

Referenz-Whisky: Auchentoshan Three Wood

◯ Kenne ich nicht

	-2	-1	0	1	2	weiß nicht
The Dalmore Cigar Malt	◯	◯	◯	◯	◯	◯
Berry's Own Selection Inchgower 1980	◯	◯	◯	◯	◯	◯
The Dalmore Cabernet Sauvignon 1973	◯	◯	◯	◯	◯	◯
The Glenronach Allardice 18 Years Old	◯	◯	◯	◯	◯	◯
The Dalmore 21 Years Old	◯	◯	◯	◯	◯	◯

Klassifizierungen

Auf Basis der ermittelten Whisky-Distanzen habe ich auch eine automatische Klassifizierung durchgeführt. Im folgenden wird jeweils eine Auswahl von Whiskys aus einer der Ermittelten Kategorien gezeigt. Teilweise ist die Auswahl zufällig getroffen worden, teilweise nach Abstand zum „Mittelpunkt“ (am nächsten bzw. am weitesten entfernte Whiskys) der Kategorie getroffen worden. Bitte kreuzen Sie zu jeder Liste an, wie nachvollziehbar sie Ihnen erscheint. Sie können außerdem ankreuzen, welche Whiskys Sie aus der Liste streichen würden um sie nachvollziehbarer zu machen.

Auswahl #1

	ausschließen
Old Pulteney 1990 Vintage	◯
Highland Harvest Organic Scotch Whisky	◯
Glenfarclas 1981 Plain Hogshead	◯
AnCnoc 1994	◯
Private Collection Imperial 1991 Port Wood Finish	◯
Glen Spey 1988 21 Years Old	◯
Magilligan 1991 Sherry Wood Finish	◯
Seagram's VO Gold 8 Years Old	◯
291 Colorado Whiskey E Batch #3	◯
Miyagikyo No Age	◯

Bewertung

◯ -2 (überhaupt nicht nachvollziehbar)

◯ -1

◯ 0

◯ 1

◯ 2 (sehr nachvollziehbar)

◯ weiß nicht / zu unbekannt

Auswahl #2

	ausschließen
Yamazaki 12 Years Old	◯
Aberlour 15 Years Old, Double Cask	◯
Auchentoshan 21 Years Old	◯
Hokuto 12 Years Old	◯
Glenkinchie 12 Years Old	◯
Glengoyne 12 Years Old	◯
Matisse 12 Years Old	◯
Black Mountain BM No2	◯
Hakushu 12 Years Old	◯
Linkwood 12 Years Old	◯

Bewertung

- ◯ -2 (überhaupt nicht nachvollziehbar)
- ◯ -1
- ◯ 0
- ◯ 1
- ◯ 2 (sehr nachvollziehbar)
- ◯ weiß nicht / zu unbekannt

Auswahl #3

	ausschließen
Longrow 10 Years Old	◯
Signatory Rosebank 1989	◯
Lombard Bowmore 1989	◯
Macleod's 8 Years Old Highland Single Malt	◯
Ardbeg 1975	◯
Box Single Malt The Explorer	◯
Talisker The Distiller's Edition Amoroso Sherry	◯
Whisky Fair Glen Scotia 'Heavily peated'	◯
The Macallan 50 Years Old, Millennium Decanter	◯
Lagavulin Distillers Edition 1991	◯

Bewertung

◯ -2 (überhaupt nicht nachvollziehbar)

◯ -1

◯ 0

◯ 1

◯ 2 (sehr nachvollziehbar)

◯ weiß nicht / zu unbekannt

Auswahl #4

	ausschließen
Four Roses 2009 Limited Edition Single Barrel	◯
The Macallan Easter Elchiles Cask Selecttion 12 Years Old	◯
Craigellachie Hotel Glenfarclas 2001 Single Cask Bottling	◯
Springbank Cask Strength 12 Years Old	◯
Abbey Whisky Ben Nevis, 16 Years Old	◯
Cooper's Choice Glenlivet 1972, 30 Years Old	◯
Chivas Brothers Glenallachie 15 Years Old, Cask Strength Edition	◯
Whyte & Mackay 13 Years Old	◯
Glenfarclas The Family Casks 1965 Release X	◯
Old Malt Cask Laphroaig 17 Years Old, Rum Finish	◯

Bewertung

◯ -2 (überhaupt nicht nachvollziehbar)

◯ -1

◯ 0

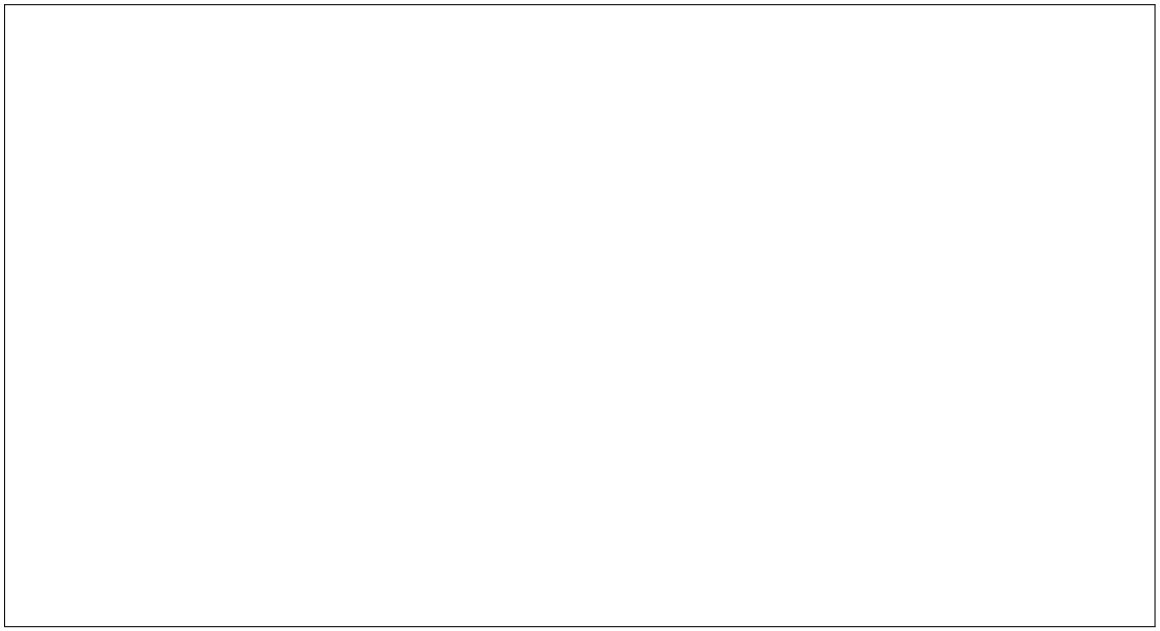
◯ 1

◯ 2 (sehr nachvollziehbar)

◯ weiß nicht / zu unbekannt

Eigene Kommentare

Zum Abschluss haben Sie hier die Möglichkeit, Ihre eigenen Kommentare und Gedanken zu äußern.



Vielen Dank für Ihre Teilnahme!

B.2. Englische Variante

Survey on Whisky Recommendations and Classifications

Being asked for a whisky recommendation always puts the respondent in a challenging situation. Whiskey is a large topic with a broad variety of expressions. Therefore, a software tool would be very helpful, which can make reliable recommendations.

I am a computer science student at Hamburg University of Applied Sciences. As part of my master's thesis, I generate whisky recommendations by comparing about 20,000 tasting notes using machine-learning methods.

In order to evaluate the results of my work, I need the opinions of some connoisseurs and experts, how comprehensible my generated recommendations are. This is not about whether you would give these recommendations in the same way, but only if you would describe these recommendations as comprehensible.

The survey takes about 15 minutes.

Level of knowledge

In order to evaluate the survey results with respect to your level of knowledge, I first need your assessment of it.

How would you rate your personal level of whisky knowledge on a scale of 1 to 7 - where 1 means beginner and 7 means expert?

	1	2	3	4	5	6	7	
Beginner	◯	◯	◯	◯	◯	◯	◯	Expert

How many years have you been dealing with whisky?

0-5	5-10	10-15	15-20	more than 20
◯	◯	◯	◯	◯

In which way do you deal with whisky?

- ◯ Hobby
- ◯ Seller
- ◯ Critic

Your data will be processed anonymously. However, if you would like to be mentioned by name in the acknowledgement of my thesis, you can enter it here.

Recommendations

Below you will find a reference whisky and its recommendations. Imagine the reference whisky being a whisky requested by a customer asking for similar whiskies. For each recommendation, please mark the extent to which you would describe it as comprehensible. I tried to use the best known whiskies from the main existing categories as reference whiskies while not picking too many of them. For each reference whisky the five best matches by calculation are shown. If you do not know the reference whisky, please mark the according field.

The evaluation scale ranges from -2 (not comprehensible at all) to 2 (very comprehensible).

Reference Whisky: Lagavulin 16 Years Old

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Laphroaig 10 Years Old, Cask Strength	◯	◯	◯	◯	◯	◯
Lagavulin 12 Years Old	◯	◯	◯	◯	◯	◯
Laphroaig 10 Years Old	◯	◯	◯	◯	◯	◯
Ardbeg 10 Years Old	◯	◯	◯	◯	◯	◯
Laphroaig Quarter Cask	◯	◯	◯	◯	◯	◯

Reference Whisky: Glenmorangie 10 Years Old

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Yamazaki 12 Years Old	◯	◯	◯	◯	◯	◯
The Glenlivet 12 Years Old	◯	◯	◯	◯	◯	◯
Black Mountain BM No2	◯	◯	◯	◯	◯	◯
Dewar's Signature	◯	◯	◯	◯	◯	◯
Midleton Single Cask 1996	◯	◯	◯	◯	◯	◯

Reference Whisky: The Balvenie 12 Years Old, Double Wood

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Speyburn 21 Years Old	◯	◯	◯	◯	◯	◯
Balblair 2004 Sherry Matured	◯	◯	◯	◯	◯	◯
Old Pulteney 1990 Vintage	◯	◯	◯	◯	◯	◯
Glen Moray 16 Years Old, Chenin Blanc Finish	◯	◯	◯	◯	◯	◯
Aberlour 1976	◯	◯	◯	◯	◯	◯

Reference Whisky: Jameson Irish Whiskey

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Redbreast 12 Years Old	◯	◯	◯	◯	◯	◯
Glen Moray 16 Years Old, Chenin Blanc Finish	◯	◯	◯	◯	◯	◯
Speyburn 21 Years Old	◯	◯	◯	◯	◯	◯
Tullamore Dew 12 Years Old	◯	◯	◯	◯	◯	◯
The Wee Dram Balblair 12 Years Old	◯	◯	◯	◯	◯	◯

Reference Whisky: Elijah Craig 12 Years Old

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Eagle Rare 17 Years Old	◯	◯	◯	◯	◯	◯
Talisker 30 Years Old	◯	◯	◯	◯	◯	◯
George T. Stagg	◯	◯	◯	◯	◯	◯
Old Forester Birthday Bourbon	◯	◯	◯	◯	◯	◯
Yamazaki 18 Years Old	◯	◯	◯	◯	◯	◯

Reference Whisky: Talisker 10 Years Old

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
Talisker 25 Years Old	◯	◯	◯	◯	◯	◯
Brora 30 Years Old	◯	◯	◯	◯	◯	◯
Lagavulin 12 Years Old	◯	◯	◯	◯	◯	◯
Caol Ila Cask Strength	◯	◯	◯	◯	◯	◯
Black Bottle 10 Years Old	◯	◯	◯	◯	◯	◯

Reference Whisky: Nikka From the Barrel

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
The Dalmore Cigar Malt	◯	◯	◯	◯	◯	◯
Tamdhu Batch Strength 002	◯	◯	◯	◯	◯	◯
Celtique Connexion Vin de Paille du Jura Wood Finish 1994	◯	◯	◯	◯	◯	◯
Girvan Single Grain 1964	◯	◯	◯	◯	◯	◯
Tomatin 14 Years Old 2002	◯	◯	◯	◯	◯	◯

Reference Whisky: Auchentoshan Three Wood

◯ I don't know the reference whisky

	-2	-1	0	1	2	don't know
The Dalmore Cigar Malt	◯	◯	◯	◯	◯	◯
Berry's Own Selection Inchgower 1980	◯	◯	◯	◯	◯	◯
The Dalmore Cabernet Sauvignon 1973	◯	◯	◯	◯	◯	◯
The Glenronach Allardice 18 Years Old	◯	◯	◯	◯	◯	◯
The Dalmore 21 Years Old	◯	◯	◯	◯	◯	◯

Classifications

I also performed an automatic classification based on the calculated whisky distances. Below you will find selections of whiskies from the categories determined. The selections were made partly by chance, partly by distance to the "centre"(closest or furthest away whiskies) of the category. Please mark for each list how comprehensible it appears to you. You can also check which whiskies you would delete from the list to make them more comprehensible.

Selection #1

	exclude
Old Pulteney 1990 Vintage	◯
Highland Harvest Organic Scotch Whisky	◯
Glenfarclas 1981 Plain Hogshead	◯
AnCnoc 1994	◯
Private Collection Imperial 1991 Port Wood Finish	◯
Glen Spey 1988 21 Years Old	◯
Magilligan 1991 Sherry Wood Finish	◯
Seagram's VO Gold 8 Years Old	◯
291 Colorado Whiskey E Batch #3	◯
Miyagikyo No Age	◯

Your rating

◯ -2 (not comprehensible at all)

◯ -1

◯ 0

◯ 1

◯ 2 (very comprehensible)

◯ don't know

Selection #2

	exclude
Yamazaki 12 Years Old	◯
Aberlour 15 Years Old, Double Cask	◯
Auchentoshan 21 Years Old	◯
Hokuto 12 Years Old	◯
Glenkinchie 12 Years Old	◯
Glengoyne 12 Years Old	◯
Matisse 12 Years Old	◯
Black Mountain BM No2	◯
Hakushu 12 Years Old	◯
Linkwood 12 Years Old	◯

Your rating

◯ -2 (not comprehensible at all)

◯ -1

◯ 0

◯ 1

◯ 2 (very comprehensible)

◯ don’t know

Selection #3

	exclude
Longrow 10 Years Old	◯
Signatory Rosebank 1989	◯
Lombard Bowmore 1989	◯
Macleod's 8 Years Old Highland Single Malt	◯
Ardbeg 1975	◯
Box Single Malt The Explorer	◯
Talisker The Distiller's Edition Amoroso Sherry	◯
Whisky Fair Glen Scotia 'Heavily peated'	◯
The Macallan 50 Years Old, Millennium Decanter	◯
Lagavulin Distillers Edition 1991	◯

Your rating

◯ -2 (not comprehensible at all)

◯ -1

◯ 0

◯ 1

◯ 2 (very comprehensible)

◯ don't know

Selection #4

	exclude
Four Roses 2009 Limited Edition Single Barrel	◯
The Macallan Easter Elchiles Cask Selecttion 12 Years Old	◯
Craigellachie Hotel Glenfarclas 2001 Single Cask Bottling	◯
Springbank Cask Strength 12 Years Old	◯
Abbey Whisky Ben Nevis, 16 Years Old	◯
Cooper's Choice Glenlivet 1972, 30 Years Old	◯
Chivas Brothers Glenallachie 15 Years Old, Cask Strength Edition	◯
Whyte & Mackay 13 Years Old	◯
Glenfarclas The Family Casks 1965 Release X	◯
Old Malt Cask Laphroaig 17 Years Old, Rum Finish	◯

Your rating

◯ -2 (not comprehensible at all)

◯ -1

◯ 0

◯ 1

◯ 2 (very comprehensible)

◯ don't know

Own Comments

Finally, you can express your own comments and thoughts here.

Thank you very much for your participation!

B.3. Befragungsergebnisse

B.3.1. Einschätzung der eigenen Expertise

Expertise	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8
Kenntnisstand (1-7)	5	5	6	5	3	3	2	5
Jahre der Beschäftigung	10 - 15	10 - 15	15 - 20	5 - 10	10 - 15	5 - 10	5 - 10	5 - 10
Art der Beschäftigung	1, 2	1, 2	2	2	1	1	1	1

B.3.2. Einschätzung der Empfehlungen

Empfehlung	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8
Lagavulin 16 Years Old								
Laphroaig 10 Years Old, Cask Strength	x	-2	0	-2	1	1	x	2
Lagavulin 12 Years Old	1	0	1	1	1	2	x	1
Laphroaig 10 Years Old	1	-2	0	-1	0	0	x	1
Ardbeg 10 Years Old	0	-1	0	-1	-1	-1	x	2
Laphroaig Quarter Cask	-1	-1	0	-2	x	x	x	2
Glenmorangie 10 Years Old								
Yamazaki 12 Years Old	x	1	2	1	-1	-1	-2	2
The Glenlivet 12 Years Old	-2	0	2	0	1	x	0	x
Black Mountain BM No2	x	x	x	x	x	x	x	x
Dewar's Signature	x	x	0	-2	x	x	x	x
Midleton Single Cask 1996	x	x	1	-1	x	x	x	x
The Balvenie 12 Years Old, Double Wood								
Speyburn 21 Years Old	0	0	1	-2	1	x	x	2
Balblair 2004 Sherry Matured	2	1	1	1	x	x	x	1
Old Pulteney 1990 Vintage	0	x	-1	-2	x	x	x	1
Glen Moray 16 Years Old, Chenin Blanc Finish	x	x	0	-2	0	x	x	x
Aberlour 1976	x	x	1	-2	x	x	x	2

Jameson Irish Whiskey								
Redbreast 12 Years Old	-1	0	2	-2	1	-1	x	1
Glen Moray 16 Years Old Chenin Blanc Finish	x	x	1	-2	1	x	x	x
Speyburn 21 Years Old	-2	0	1	-2	0	x	x	x
Tullamore Dew 12 Years Old	2	1	1	-1	1	1	0	1
The Wee Dram Balblair 12 Years Old	-2	x	x	x	x	x	x	x
Elijah Craig 12 Years Old								
Eagle Rare 17 Years Old	1	-2	2	1	x	x	x	x
Talisker 30 Yeas Old	-1	-2	-2	-2	x	x	x	x
George T. Stagg	1	x	1	0	x	x	x	x
Old Forester Birthday Bourbon	x	x	x	1	x	x	x	x
Yamazaki 18 Years Old	x	-2	-2	-2	x	x	x	x
Talisker 10 Years Old								
Talisker 25 Years Old	1	1	-1	1	1	2	1	2
Brora 30 Years Old	x	x	-1	-2	x	x	x	x
Lagavulin 12 Years Old	1	0	2	0	2	1	x	1
Caol Ila Cask Strength	1	0	2	-1	2	2	x	1
Black Bottle 10 Years Old	-1	-2	1	x	x	x	x	x
Nikka From the Barrel								
The Dalmore Cigar Malt	2	-2	-1	-2	x	x	x	1
Tamdhu Batch Strength 002	2	-2	-2	-2	x	x	x	x
Celtique Connexion Vin de Paille du Jura Wood Finish 1994	x	x	x	x	x	x	x	x
Girvan Single Grain 1964	x	0	1	-2	x	x	x	x
Tomatin 14 Years Old 2002	1	x	0	-2	x	x	x	x
Auchentoshan Three Wood								
The Dalmore Cigar Malt	0	-1	2	0	x	x	x	x

Berry's Own Selection Inchgower 1980	x	x	x	-2	x	x	x	x
The Dalmore Cabernet Sauvignon 1973	1	x	x	-2	x	x	x	x
The Glendronach Allardice 18 Years Old	x	-2	0	-2	x	x	x	x
The Dalmore 21 Years Old	0	x	2	-1	x	x	x	x

B.3.3. Einschätzung der Cluster

Auswahl	**Exp1**	**Exp2**	**Exp3**	**Exp4**	**Exp5**	**Exp6**	**Exp7**	**Exp8**
Auswahl #1	x	1	x	-2	x	x	x	0
Auswahl #2	-2	2	2	-1	1	x	x	3
Auswahl #3	-2	1	2	-2	x	x	x	1
Auswahl #4	-2	-1	-2	-2	0	x	x	0

Literaturverzeichnis

[Arthur und Vassilvitskii 2007] Arthur, David ; Vassilvitskii, Sergei: k-means++: The advantages of careful seeding. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* Society for Industrial and Applied Mathematics (Veranst.), 2007, S. 1027–1035

[Bathgate 2003] Bathgate, George N.: History of the development of whiskey distillation. In: Russell, Inge (Hrsg.) ; Bamforth, Charles W. (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. First edition. San Diego : Academic Press, 2003 (Handbook of Alcoholic Beverages), S. 1 – 24. – URL http://www.sciencedirect.com/science/article/pii/B978012669202050018X. – ISBN 978-0-12-669202-0

[Behnel u. a. 2018] Behnel, Stefan ; Faassen, Martijn ; Bicking, Ian: *lxml: XML and HTML with Python*. 2018. – URL https://lxml.de/. – Zugriffsdatum: 27.07.2018

[Bird und Loper 2004] Bird, Steven ; Loper, Edward: NLTK: The Natural Language Toolkit. In: *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2004 (ACLdemo '04). – URL http://dx.doi.org/10.3115/1219044.1219075

[Bojanowski u. a. 2016] Bojanowski, Piotr ; Grave, Edouard ; Joulin, Armand ; Mikolov, Tomas: Enriching Word Vectors with Subword Information. In: *CoRR* abs/1607.04606 (2016). – URL http://arxiv.org/abs/1607.04606

[Boose 1989] Boose, John H.: A survey of knowledge acquisition techniques and tools. In: *Knowledge acquisition* 1 (1989), Nr. 1, S. 3–37

[Bringhurst u. a. 2014] Bringhurst, Tom A. ; Brosnan, Inge ; Stewart, Graham G.: Water: an essential raw material for whisk(e)y production. In: Russell, Inge (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2014, S. 291

– 298. – URL http://www.sciencedirect.com/science/article/pii/B9780124017351000167. – ISBN 978-0-12-401735-1

[Bringhurst und Brosnan 2014] Bringhurst, Tom A. ; Brosnan, James: Scotch whisky: raw material selection and processing. In: Russell, Inge (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2014, S. 49 – 122. – URL http://www.sciencedirect.com/science/article/pii/B9780124017351000064. – ISBN 978-0-12-401735-1

[Brossard 2018] Brossard, Patrick: *Whisky News*. 2018. – URL https://www.whisky-news.com/. – Zugriffsdatum: 26.07.2018

[Conner 2014] Conner, John: Maturation. In: Russell, Inge (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2014, S. 199 – 220. – URL http://www.sciencedirect.com/science/article/pii/B9780124017351000118. – ISBN 978-0-12-401735-1

[Cooke 1994] Cooke, Nancy J.: Varieties of knowledge elicitation techniques. In: *International Journal of Human-Computer Studies* 41 (1994), Nr. 6, S. 801–849

[CouchDB 2018] CouchDB: *Apache CouchDB*. 2018. – URL http://couchdb.apache.org. – Zugriffsdatum: 30.08.2018

[Distiller 2018] Distiller: *Distiller*. 2018. – URL https://distiller.com. – Zugriffsdatum: 14.08.2018

[Dolan 2003] Dolan, Timothy C.: Malt whiskies: raw materials and processing. In: Russell, Inge (Hrsg.) ; Bamforth, Charles W. (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2003 (Handbook of Alcoholic Beverages), S. 27 – 73. – URL http://www.sciencedirect.com/science/article/pii/B9780126692020500191. – ISBN 978-0-12-669202-0

[Ester u. a. 1996] Ester, Martin ; Kriegel, Hans-Peter ; Sander, Jörg ; Xu, Xiaowei u. a.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd* Bd. 96, 1996, S. 226–231

[Fayyad u. a. 1996a] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: From data mining to knowledge discovery in databases. In: *AI magazine* 17 (1996), Nr. 3, S. 37. – URL http://dx.doi.org/10.1609/aimag.v17i3.1230

[Fayyad u. a. 1996b] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Commun. ACM* 39 (1996), nov, Nr. 11, S. 27–34. – URL http://doi.acm.org/10.1145/240455.240464. – ISSN 0001-0782

[Garg u. a. 2017] Garg, Neelansh ; Sethupathy, Apuroop ; Tuwani, Rudraksh ; Dokania, Shubham ; Iyer, Arvind ; Gupta, Ayushi ; Agrawal, Shubhra ; Singh, Navjot ; Shukla, Shubham ; Kathuria, Kriti u. a.: FlavorDB: a database of flavor molecules. In: *Nucleic acids research* 46 (2017), Nr. D1, S. D1210–D1216

[Glazman u. a. 2018] Glazman, Daniel ; Linss, Peter ; Çelik, Tantek ; Williams, John ; Hickson, Ian ; Etemad, Elika: Selectors Level 3 / W3C. W3C, January 2018. – W3C Candidate Recommendation. – URL https://www.w3.org/TR/2018/CR-selectors-3-20180130/

[Hunter 2007] Hunter, J. D.: Matplotlib: A 2D graphics environment. In: *Computing In Science & Engineering* 9 (2007), Nr. 3, S. 90–95

[Jack 2014] Jack, Frances R.: Sensory analysis. In: Russell, Inge (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2014, S. 229 – 242. – URL http://www.sciencedirect.com/science/article/pii/B9780124017351000131. – ISBN 978-0-12-401735-1

[Jack und Steele 2002] Jack, Frances R. ; Steele, Gordon M.: Modelling the sensory characteristics of Scotch whisky using neural networks—a novel tool for generic protection. In: *Food Quality and Preference* 13 (2002), Nr. 3, S. 163–172

[Jackson 2017] Jackson, Michael: *Whisky: The definitive world guide*. Dorling Kindersley Ltd, 2017

[Jain u. a. 1999] Jain, A. K. ; Murty, M. N. ; Flynn, P. J.: Data Clustering: A Review. In: *ACM Comput. Surv.* 31 (1999), sep, Nr. 3, S. 264–323. – URL http://doi.acm.org/10.1145/331499.331504. – ISSN 0360-0300

[Karam 2017] Karam, Ramzi: *Using Word2vec for Music Recommendations*. 2017. – URL https://towardsdatascience.com/using-word2vec-for-music-recommendations-bb9649ac2484. – Zugriffsdatum: 01.08.2018

[Klaverstijn 2018] Klaverstijn, Thijs: *Words of Whisky*. 2018. – URL https://wordsofwhisky.com/. – Zugriffsdatum: 26.07.2018

[Kosinus-Ähnlichkeit 2018] Kosinus-Ähnlichkeit: *Kosinus-Ähnlichkeit — Wikipedia, Die freie Enzyklopädie*. 2018. – URL https://de.wikipedia.org/wiki/Kosinus-%C3%84hnlichkeit. – Zugriffsdatum: 29.07.2018

[Krzus 2018] Krzus, Matt: *Whiskey Embeddings*. 2018. – URL http://wrec.herokuapp.com/methodology. – Zugriffsdatum: 05.08.2018

[Lardin 2018] Lardin, Miguel Angel B.: *A Wardrobe of Whisky*. 2018. – URL http://www.awardrobeofwhisky.com. – Zugriffsdatum: 27.07.2018

[Maaten und Hinton 2008] Maaten, Laurens van d. ; Hinton, Geoffrey: Visualizing data using t-SNE. In: *Journal of Machine Learning Research* 9 (2008), nov, S. 2579–2605

[Malt 2018] Malt: *Malt*. 2018. – URL https://malt-review.com/. – Zugriffsdatum: 26.07.2018

[Master Of Malt 2018] Master Of Malt: *Master of Malt*. 2018. – URL https://www.masterofmalt.com. – Zugriffsdatum: 26.07.2018

[McCormick 2016] McCormick, C.: *Word2Vec Tutorial - The Skip-Gram Model*. 2016. – URL http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/. – Zugriffsdatum: 01.08.2018

[McCormick 2018] McCormick, C.: *Applying word2vec to Recommenders and Advertising*. 2018. – URL http://mccormickml.com/2018/06/15/applying-word2vec-to-recommenders-and-advertising/. – Zugriffsdatum: 01.08.2018

[Mikolov 2013] Mikolov, T.: *de-obfuscated Python + question*. 2013. – URL https://groups.google.com/d/msg/word2vec-toolkit/NLvYXU99cAM/E5ld8LcDxlAJ. – Zugriffsdatum: 03.08.2018. – Google-Groups-Konversation

[Mikolov u. a. 2013a] Mikolov, Tomas ; Chen, Kai ; Corrado, Greg ; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. In: *CoRR* abs/1301.3781 (2013). – URL http://arxiv.org/abs/1301.3781

[Mikolov u. a. 2013b] Mikolov, Tomas ; Sutskever, Ilya ; Chen, Kai ; Corrado, Greg S. ; Dean, Jeff: Distributed Representations of Words and Phrases and their Compositionality. In: Burges, C. J. C. (Hrsg.) ; Bottou, L. (Hrsg.) ; Welling, M. (Hrsg.) ; Ghahramani, Z. (Hrsg.) ; Weinberger, K. Q. (Hrsg.): *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, S. 3111–3119. – URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

[Miller 1995] Miller, George A.: WordNet: a lexical database for English. In: *Communications of the ACM* 38 (1995), Nr. 11, S. 39–41

[Murray 2016] Murray, Jim: *Jim Murray's Whisky Bible 2017*. Dram Good Books, 2016

[MySQL 2018] MySQL: *MySQL*. 2018. – URL https://www.mysql.com/. – Zugriffsdatum: 27.07.2018

[Pedregosa u. a. 2011] Pedregosa, F. ; Varoquaux, G. ; Gramfort, A. ; Michel, V. ; Thirion, B. ; Grisel, O. ; Blondel, M. ; Prettenhofer, P. ; Weiss, R. ; Dubourg, V. ; Vanderplas, J. ; Passos, A. ; Cournapeau, D. ; Brucher, M. ; Perrot, M. ; Duchesnay, E.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830

[Pennington u. a. 2014] Pennington, Jeffrey ; Socher, Richard ; Manning, Christopher: GloVe: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, S. 1532–1543

[Piggott und Jardine 1979] Piggott, JR ; Jardine, SP: Descriptive sensory analysis of whisky flavour. In: *Journal of the Institute of Brewing* 85 (1979), Nr. 2, S. 82–85

[Reddit 2018] Reddit: *Reddit*. 2018. – URL https://www.reddit.com. – Zugriffsdatum: 27.07.2018

[Regulation (EC) No 110/2008 2008] : *Regulation (EC) No 110/2008 of the European Parliament and of the Council of 15 January 2008 on the definition, description, presentation, labelling and the protection of geographical indications of spirit drinks and repealing Council Regulation (EEC) No 1576/89*. 2008. – URL https://eur-lex.europa.eu/eli/reg/2008/110(1)/oj

[Řehůřek und Sojka 2010] Řehůřek, Radim ; Sojka, Petr: Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta : ELRA, may 2010, S. 45–50. – http://is.muni.cz/publication/884893/en

[Ricci u. a. 2011] Ricci, Francesco ; Rokach, Lior ; Shapira, Bracha: *Introduction to Recommender Systems Handbook*. S. 1–35. In: Ricci, Francesco (Hrsg.) ; Rokach, Lior (Hrsg.) ; Shapira, Bracha (Hrsg.) ; Kantor, B. P. (Hrsg.): *Recommender Systems Handbook"*. Boston, MA : Springer US, 2011. – URL http://dx.doi.org/10.1007/978-0-387-85820-3_1. – ISBN 978-0-387-85820-3

[Robie u. a. 2017] Robie, Jonathan ; Spiegel, Josh ; Dyck, Michael: XML Path Language (XPath) 3.1 / W3C. W3C, March 2017. – W3C Recommendation. – URL https://www.w3.org/TR/2017/REC-xpath-31-20170321/

[Ronde 2016] Ronde, Ingvar: *Malt Whisky Yearbook 2017*. MapDig Media Limited, 2016

[Schakel und Wilson 2015] Schakel, Adriaan M. J. ; Wilson, Benjamin J.: Measuring Word Significance using Distributed Representations of Words. In: *CoRR* abs/1508.02297 (2015). – URL http://arxiv.org/abs/1508.02297

[Schole 2017a] Schole, Joachim: *Whisky-Empfehlungen*, Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung Hauptseminar, März 2017. – http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2016-hsem/schole/bericht.pdf

[Schole 2017b] Schole, Joachim: *Whisky-Empfehlungen - Aufbau eines Datenkorpus als Grundlage weiterer Experimente zur Ermittlung von Distanzen zwischen Whiskys*, Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung Grundprojekt, April 2017. – http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2017-proj/schole.pdf

[Schole 2018] Schole, Joachim: *Whisky-Empfehlungen - Evaluierung eines KDD-Prozesses zur Entwicklung eines Empfehlungssystems*, Hochschule für angewandte Wissenschaften Hamburg, Ausarbeitung Hauptprojekt, February 2018. – http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2018-proj/schole.pdf

[Schütze u. a. 2008] Schütze, Hinrich ; Manning, Christopher D. ; Raghavan, Prabhakar: *Introduction to information retrieval*. Bd. 39. Cambridge University Press, 2008

[Scotch Whisky Act 1988] *Scotch Whisky Act*. 1988. – URL http://www.legislation.gov.uk/ukpga/1988/22. – Chapter 22

[Scotch Whisky Regulations 2009] *The Scotch Whisky Regulations*. 2009. – URL http://www.legislation.gov.uk/uksi/2009/2890. – SI 2009/2890

[Scotchwhisky.com 2018] Scotchwhisky.com: *Scotchwhisky.com*. 2018. – URL https://scotchwhisky.com. – Zugriffsdatum: 04.08.2018

[Selenium 2018] Selenium: *Selenium*. 2018. – URL http://www.seleniumhq.org. – Zugriffsdatum: 29.07.2018

[Sharafi 2013] Sharafi, Armin: *Knowledge Discovery in Databases*. S. 51–108. In: *Knowledge Discovery in Databases: Eine Analyse des Änderungsmanagements in der Produktentwicklung*. Wiesbaden : Springer Fachmedien Wiesbaden, 2013. – URL http://dx.doi.org/10.1007/978-3-658-02002-6_3. – ISBN 978-3-658-02002-6

[Silhouettenkoeffizient 2018] Silhouettenkoeffizient: *Silhouettenkoeffizient — Wikipedia, Die freie Enzyklopädie*. 2018. – URL https://de.wikipedia.org/wiki/Silhouettenkoeffizient. – Zugriffsdatum: 14.08.2018

[Stewart u. a. 2014] Stewart, Graham ; Russell, Inge ; Anstruther, Anne: An introduction to whisk(e)y and the development of Scotch whisky. In: Russell, Inge (Hrsg.) ; Stewart, Graham G. (Hrsg.): *Whisky*. Second edition. San Diego : Academic Press, 2014, S. 1 – 5. – URL http://www.sciencedirect.com/science/article/pii/B9780124017351000015. – ISBN 978-0-12-401735-1

[Thomson 2018] Thomson, Tom: *Toms Whisky Reviews*. 2018. – URL http://www.tomswhiskyreviews.com/. – Zugriffsdatum: 26.07.2018

[Whiskey Reviewer 2018] Whiskey Reviewer: *The Whiskey Reviewer*. 2018. – URL http://whiskeyreviewer.com/. – Zugriffsdatum: 26.07.2018

[Whisky Advocate 2018] Whisky Advocate: *Whisky Advocate*. 2018. – URL http://whiskyadvocate.com. – Zugriffsdatum: 17.07.2018

[Whisky Intelligence 2018] Whisky Intelligence: *Whiskyintelligence.com*. 2018. – URL http://whiskyintelligence.com. – Zugriffsdatum: 04.08.2018

[Whisky Magazine 2018] Whisky Magazine: *Whisky Magazine*. 2018. – URL https://www.whiskymag.com. – Zugriffsdatum: 30.08.2018

[Whisky Monitor 2018] Whisky Monitor: *Whisky Monitor*. 2018. – URL https://www.whisky-monitor.com. – Zugriffsdatum: 04.08.2018

[Whisky Notes 2018] Whisky Notes: *Whisky Notes*. 2018. – URL https://www.whiskynotes.be/. – Zugriffsdatum: 26.07.2018

[Whisky.de 2018] Whisky.de: *Whisky.de*. 2018. – URL http://www.whisky.de. – Zugriffsdatum: 05.08.2018

[Whiskyology 2018] Whiskyology: *Whiskyology*. 2018. – URL https://www.whiskyology.com. – Zugriffsdatum: 26.07.2018

[Wishart 2009] Wishart, David: The flavour of whisky. In: *Significance* 6 (2009), Nr. 1, S. 20–26

[Wishart 2014] Wishart, D.S.: *FooDB: the food database. FooDB version 1.0*. 2014. – URL http://foodb.ca/. – Zugriffsdatum: 29.07.2018

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 31.08.2018 Joachim Schole