Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Grundprojekt

Joachim Schole

Whisky-Empfehlungen

Aufbau eines Datenkorpus als Grundlage weiterer Experimente zur Ermittlung von Distanzen zwischen Whiskys

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Joachim Schole

Whisky-Empfehlungen

Aufbau eines Datenkorpus als Grundlage weiterer Experimente zur Ermittlung von Distanzen zwischen Whiskys

Ausarbeitung Grundprojekt eingereicht im Rahmen des Grundprojekt

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

Eingereicht am: 2. April 2017

Inhaltsverzeichnis

1 Einleitung

Die Frage nach einer Whisky-Empfehlung stellt den gefragten vor mehrere Probleme. Für eine akkurate Einschätzung muss er nicht nur die Geschmäcker möglichst vieler verschiedener Whiskys kennen - und diese bestenfalls selbst probiert haben - , sondern auch den Geschmack des Fragenden kennen und dabei die eigene, in der Regel subjektive Meinung ausblenden. Angesichts der Menge an verschiedenen Ausprägungen nach Region und Destillerie erscheint es unmöglich, ausreichend Whiskys zu kennen oder selbst bei einem großen Kenntnisstand alle bekannten Whiskys zu berücksichtigen. Hinzu kommt, dass nicht jeder Barkeeper sich mit dem gesamten Angebot an seinem Arbeitsplatz auskennen kann oder möchte. Der gefragte hätte in diesem Fall durch ein Empfehlungssystem die Möglichkeit, eine qualifizierte und objektive Empfehlung anhand genannter Präferenzen zu geben. Auch ist es denkbar, dass eine Bar den Lieblingswhisky eines Kunden nicht anbietet oder vorrätig hat. In diesem Fall kann mit einem geeigneten System ein möglichst ähnlicher Whisky unter den vorrätigen ermittelt werden oder je nach Aufbau des Systems ein Whisky, der sich vom gewünschten unterscheidet, bekanntermaßen aber vielen anderen Fans dieses Whiskys gefällt.

Hier setzt diese Arbeit an. Ziel soll es sein, die Grundlage für ein Whisky-Empfehlungssystem zu entwickeln. Dies soll durch einen auf das Problem zugeschnittenen KDD-Prozess geschehen, wobei sogenannte Tasting Notes die Datengrundlage bilden. Dabei stellen sich einige Probleme. Zunächst muss ein ausreichend tiefes Verständnis der Whiskyvielfalt beziehungsweise des Whiskygeschmacks vorhanden sein, um eine geeignete Datenrepräsentation für die weitere Verarbeitung zu entwickeln. Weiter muss anhand dieser Repräsentationsform ein Distanzmaß für die Geschmäcker zweier Whiskys festgelegt werden. Über diese Distanzen können bereits die Nachbarn eines Whiskys ermittelt werden. Mittels Clustering können daraufhin Kategorien beziehungsweise Gruppen ermittelt werden. Diese bieten dann die Möglichkeit, weitere Kenntnisse zu erlangen. Eine naheliegende Frage ist die, ob und zu welchem Grad die Herkunft eines Whiskys tatsächlich Aufschluss auf den Geschmack gibt.

2 Die Domäne Whisky

Dieses Kapitel gibt einen Einblick in die Domäne Whisky und in die für diese Arbeit bedeutenden Eigenschaften. Besonders von Interesse ist die geschmackliche Zusammensetzung und der allgemeine Aufbau von Geschmacksbeschreibungen. Die wesentlichen Aussagen des Kapitels stammen aus Jack (2014). Um ein Verständnis von Whisky zu erlangen ist es vor allem wichtig, den Herstellungsprozess und die aromatische Zusammensetzung zu verstehen. Whiskys und Destillerien unterscheiden sich im wesentlichen durch das Aroma ihrer Produkte. Produzenten legen viel Wert auf die Entwicklung eines eigenen Geschmacks, welcher sich von der Konkurrenz abhebt. Die traditionelle Whisky-Erzeugung erlaubt keine zusätzlichen Aromastoffe, weshalb alle Geschmacksnoten aus den Rohmaterialien bezogen werden oder während den einzelnen Produktionsschritten entstehen. Die Produktionsschritte und die wichtigsten dabei entstehenden Aromen lassen sich wie folgt zuordnen: Aus den Rohmaterialien wie getorftem (geräuchertem) Whiskymalz entstehen Aromen wie (medizinischer) Rauch, Malz, Biskuit und weitere. Dies ist abhängig von der Getreidesorte. Bei der Gärung entstehen je nach Hefestamm, Dauer und Temperatur als Beiprodukte verschiedene Ester, Aldehyde und Säuren. Während der Destillation werden einige vorher entstandene Aromen wieder herausgefiltert. Gleichzeitig entstehen durch die Hitze und die damit einhergehenden chemischen Reaktionen neue Aromen wie Getreide, Blumen, Gras, verbrannt und Schwefel. Während der Reifung bezieht der Whisky neben Aromen auch Farbe aus dem Fass. Zu den hier entstehenden Aromen gehören vor allem Vanillin und Eichenlacton (Kokosnussaroma). Dies ist abhängig von der Holzsorte und dem zuvor darin gereiften Produkt wie beispielsweise Sherry. Gleichzeitig verfliegen einige zuvor entstandene Aromen mit dem sogenannten *Angel's Share*. Weitere aromabildende Reaktionen finden über einen längeren Zeitraum statt, was die Dauer der Fassreifung zu einem wichtigen Faktor macht.

Neben den riechbaren Aromen spielen auch die geschmacklichen Hauptmerkmale süß, bitter und sauer eine große Rolle, wobei letzteres unerwünscht ist. Weiter spielt auch das Mundgefühl eine Rolle. Dieses unterteilt sich in die Eigenschaften (Mund-)Wärmend, Adstringenz und Vollmundig. Adstringenz ist ein Begriff aus der Weinsprache, welcher „die Fähigkeit eines Weines, ein „raues", „pelziges" Mundgefühl zu verursachen" (Wikipedia (2016)) beschreibt.

Auch visuelle Eigenschaften haben Einfluss auf die Wahrnehmung eines Whiskys. Besonders die Farbe, aber auch die Klarheit eines Whiskys spielen hierbei eine Rolle. Die Farbe darf mittels Einsatz von Karamell angepasst und die Klarheit durch Kühlfiltrierung verbessert werden. Diese Arbeit berücksichtigt die visuellen Eigenschaften zunächst nicht weiter, da Informationen hierzu nicht immer verfügbar sind und der Fokus vorerst auf dem Geschmack liegt.

Abbildung 2.1: Das Nosing-Wheel nach (Jack, 2014, S. 238)

Da verschiedene Aromen auf verschiedene Produktionsschritte zurückzuführen sind, muss die Entwicklung dieser im Laufe der Produktion immer wieder überprüft werden. Unter anderem hierzu dient die aromatische Bewertung von Whiskys. Weitere Anwendungsfälle sind das *Blending* und die Erstellung von werbekräftigen Produktbeschreibungen. Vorwiegend aufgrund des hohen Alkoholgehalts werden die Aromen von Whiskys in der Regel durch das *Nosing* (riechen) bestimmt. Da noch nicht alle Aromen einem entsprechenden Produktionsschritt zugeordnet werden können, ist der Einsatz von Messinstrumenten nicht oder nur teilweise

möglich. Zur Unterstützung beim Nosing und auch um ein einheitliches Vokabular zu schaffen, sind die weit verbreiteten *Nosing Wheels* wie beispielsweise in Abbildung 2.1 wichtig. Diese bieten einen Überblick über die verschiedenen Vokabeln beim Nosing und gleichzeitig die hierarchische Anordnung dieser in Kategorien und Unterkategorien. So lässt sich ermitteln, in welcher geschmacklichen Distanz einzelne Aromen zueinander stehen. Folglich bieten Nosing Wheels eine mögliche Datengrundlage für eine Aromen-Ontologie.

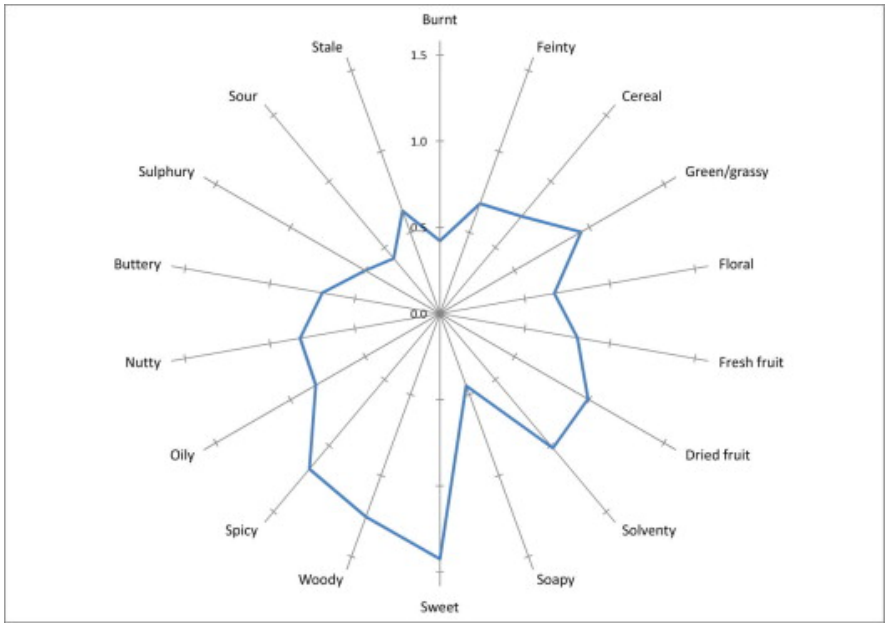

Abbildung 2.2: Beispiel für ein Flavour Profile (Jack, 2014, S. 241)

In einigen Fällen reicht die einfache qualitative Bewertung der Aromen nicht aus und eine quantitative Bewertung der Intensität der einzelnen Aromen anhand von Skalen ist notwendig. Abbildung 2.2 zeigt Beispielhaft das *Flavour Profile* eines Bourbons.

Prinzipiell sind alle verfügbaren Daten zu einzelnen Whiskys von Interesse. Für den hier geplanten ersten Ansatz sollen nur die Tasting Notes als Variablen dienen, um die reine Geschmackliche Nähe zwischen den Whiskys berechnen zu können. Von großem Vorteil wäre eine Datenquelle, welche auch quantitative Bewertungen, also Flavour Profiles bietet.

3 Technische Grundlagen

Dieses Kapitel fasst die zugrundeliegenden Techniken zusammen. Zunächst erfolgt eine Einführung in Empfehlungssysteme und daraufhin eine Beschreibung des KDD-Prozesses. Daraufhin wird näher auf das Data Mining und Clustering eingegangen. Zudem erfolgt eine Einführung in den Algorithmus *Word2vec*.

3.1 Empfehlungssysteme

Dieses Kapitel bietet eine Einführung in Empfehlungssysteme. Die wesentlichen Aussagen stammen aus Ricci u. a. (2011).

Durch die große Auswahl an Produkten im Internet sind Empfehlungssysteme heutzutage von großer Bedeutung. Ein normaler Nutzer hat beim Einkauf im Internet nicht die Zeit oder Muße, sich aus teilweise hunderten Varianten das beste oder passendste Produkt herauszusuchen. Dies ist der zentrale Anwendungsfall für Empfehlungssysteme. Ein Empfehlungssystem ist ein Software-Tool, welches einem Nutzer nützliche *Items* vorschlägt. Dabei beschreibt Item das Objekt von Interesse. In dieser Arbeit sind die Items Whiskys. In der Regel ist ein Empfehlungssystem auf Items einer speziellen Domäne ausgerichtet. Je nach System kann es sich bei den Items beispielsweise um Nachrichtenartikel, Musikstücke oder andere Verkaufsartikel handeln. Denkbar sind ebenso auf einzelne Patienten zugeschnittene Behandlungsmethoden in der Medizin. Als Hauptzielgruppe von Empfehlungssystemen gelten in der entsprechenden Domäne unerfahrene Menschen. Für erfahrene Kunden bietet ein Empfehlungssystem die Möglichkeit, bisher unbeachtete oder besonders neue Items aus der entsprechenden Kategorie zu entdecken. So ist es denkbar, dass ein Whisky-Experte, welcher sich noch nie mit deutschen Destillerien befasst hat, hier durch ein Empfehlungssystem einen leichteren Einstieg erhält.

Der Einsatz von Empfehlungssystemen ist für den Betreiber (in der Regel ein Händler) ebenso von Interesse wie für den Nutzer. Der Betreiber hat vorwiegend die Motivation, seine Verkaufszahlen durch gute Empfehlungen zu erhöhen, während der Nutzer möglichst schnell zum besten oder geeignetsten Produkt gelangen möchte. Neben diesen Motiven existieren noch einige weitere wie beispielsweise die Erhöhung der Nutzerzufriedenheit und der damit

einhergehenden -Treue und auch das Motiv des Nutzers, sich zu Informieren oder aber auch das System nach Möglichkeit durch eigene Empfehlungen und Bewertungen zu beeinflussen.

Die Grundlage eines Empfehlungssystems bilden die drei wesentlichen Datentypen Items, Nutzer und Transaktionen. Items bezeichnet die zu filternden Gegenstände oder Objekte und sämtliche dazu verfügbaren beziehungsweise benötigten Daten. Der Nutzer kann je nach System lediglich aus den Daten eines Kunden bestehen oder aber auch zusätzlich aus einer Menge an vergangenen Transaktionen. Transaktionen sind Aktionen, die zwischen dem Nutzer und einem Item stattfinden. Dazu gehören beispielsweise Bewertungen, Einkäufe und gegebenenfalls Rücksendungen oder Beschwerden.

Wie genau ein System die passendsten Items ermittelt und darstellt, variiert stark. Je nach Domäne bieten sich verschiedene Varianten an. Dabei kann zwischen sechs verschiedenen Ansätzen unterschieden werden: Inhaltsbasierte (*Content-based*) Empfehlungssysteme vergleichen Items anhand ihrer Eigenschaften. Sie empfehlen einem Nutzer solche Items, welche anderen ähneln, die er beispielsweise bereits in der Vergangenheit für gut befunden hat. Das *Collaborative Filtering* vergleicht verschiedene Nutzer anhand ihrer Präferenzen und empfiehlt einem Nutzer die Items, die andere Nutzer mit ähnlichem Geschmack bereits für gut befunden haben. Beim demographischen Ansatz werden bestimmte Eigenschaften des Nutzers wie Standort, Sprache, Alter etc. als Grundlage für Empfehlungen genutzt. Wissensbasierte Systeme bauen auf tiefes domänenspezifisches Wissen über Items und Nutzer und empfehlen einem Nutzer Items entsprechend seiner Bedürfnisse nach einem Problemlösungsprinzip. Ähnlich wie das Collaborative Filtering funktioniert der *Community-based* Ansatz. Der Unterschied liegt hier darin, dass statt Nutzern mit ähnlichen Interessen Personen aus dem sozialen Umfeld des Nutzers als Empfehlungsgrundlage fungieren. So erhält der Nutzer quasi Empfehlungen von seinen Bekannten, was oft zu mehr Vertrauen in die Empfehlungen führt. Hybride Systeme sind nach Bedarf aus den oben genannten Ansätzen kombinierbar, um die schwächen der einzelnen Methoden zu beheben.

3.2 Knowledge Discovery in Databases

Dieses Kapitel bietet einen Überblick über den KDD-Prozess und zugehörige Begriffe. Die wesentlichen Aussagen stammen, sofern nicht anders angegeben, aus den Arbeiten Fayyad u. a. (1996a), Fayyad u. a. (1996b) und Sharafi (2013).

Knowledge Discovery in Databases (KDD) beschreibt allgemein den interdisziplinären Prozess der automatisierten Wissensgewinnung aus (Roh-)Datenmengen, welche zu groß sind, um manuell ausgewertet zu werden. Rohdaten sind in der Regel erst dann von Wert, wenn das

in ihnen liegende Wissen ermittelt wird. Der KDD-Prozess ist in mehrere Schritte unterteilt. Diese sind grob in Abbildung 3.1 dargestellt.

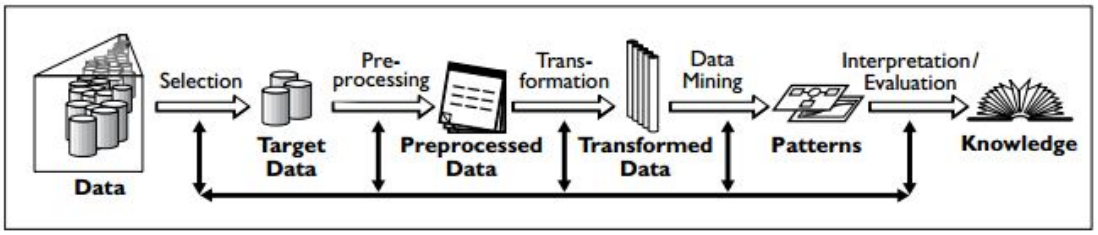

Abbildung 3.1: Der KDD-Prozess (Fayyad u. a., 1996b, S. 29)

Zunächst muss ein ausreichend tiefes Verständnis der Domäne erlangt werden, um das Ziel des KDD-Prozesses und die dazugehörigen Probleme exakt benennen zu können. Hierzu dient die Beschreibung der Domäne Whisky in Kapitel 2. Dieser erste Schritt kann einen Großteil der gesamten Arbeit einnehmen. Daraufhin kann die Auswahl geeigneter Rohdaten erfolgen. Diese müssen bereinigt und vorverarbeitet werden. Dazu gehört die Entfernung von Fehlerhaften Daten und die Rauscherkennung. Außerdem müssen Probleme auf Schema- und Instanzlevel beseitigt werden, welche sowohl bei der Verwendung einer einzigen als auch mehrerer Quellen vorkommen können. Diese umfassen beispielsweise Fehler bei der Datenerfassung und bei mehreren Quellen heterogene Datenmodelle. Hiernach findet eine Transformation der Daten in eine geeignete Repräsentationsform statt. Daraufhin muss eine Auswahl der geeigneten Data-Mining-Methoden stattfinden. Auf dieser Entscheidung basierend müssen Modelle, Algorithmen und Parameter zur Mustererkennung in den Daten ausgewählt werden. Nach diesen Entscheidungen findet die eigentliche Durchführung des Data Mining statt, woraufhin die entdeckten Patterns interpretiert und somit Wissen erlangt wird. Dieses Wissen dient als Grundlage für weitere Entscheidungen. Dazu gehört auch die Möglichkeit, den KDD-Prozess anzupassen und erneut zu durchlaufen.

3.2.1 Data Mining

Ein zentraler Bestandteil im KDD-Prozess ist das Data Mining. Ziel dieses Schritts ist das erkennen von Mustern in den aufbereiteten Rohdaten. Diese Muster dienen einem übergeordnetem Ziel, welches entweder die Verifikation aufgestellter Hypothesen oder die Entdeckung neuen Wissens ist. Die Entdeckung kann dabei entweder zum Ziel haben, zukünftiges Verhalten mit einer ermittelten Wahrscheinlichkeit vorauszusagen oder die in den Rohdaten liegenden Muster verständlich zu beschreiben.

Die verschiedenen Data-Mining-Methoden lassen sich in die Kategorien Klassenbildung (*Clustering*), Klassifizierung (*Classification*), Assoziationsanalyse und Zeitreihenanalyse unterteilen. Diese Kategorien teilen sich weiter in Unterkategorien und spezielle Algorithmen auf. Die Klassenbildung hat das automatische partitionieren von Daten zu Klassen (Clustern) zum Ziel, während die Klassifizierung Daten in vorgegebene Klassen einteilt. Die Assoziationsanalyse verfolgt das Ziel der Entdeckung von Beziehungen und Assoziationen mit dem Ziel, häufig gemeinsam auftretende Datensätze zu ermitteln.

In der Regel lassen sich Data-Mining-Algorithmen auf die drei Komponenten Muster(-repräsentation), Musterevaluation und Suchalgorithmus zurückführen. Die Musterrepräsentation bietet eine Sprache zur Beschreibung möglicher zu entdeckender Muster. Die Musterevaluation geschieht anhand von Präferenzkriterien, welche einen Vergleich der Eignung verschiedener gefundener Muster ermöglichen. Der Suchalgorithmus findet schließlich das oder die geeignetsten Muster in den Daten unter Verwendung des Musters und der Präferenzkriterien.

3.2.2 Clustering

Clustering ist eine Data-Mining-Methode und beschreibt die automatische Unterteilung von Daten in Gruppen. Dabei gibt es je nach Anwendungsfall und Domäne verschiedene Vorgehensweisen. In der Regel besteht ein Clustering-Prozess aus folgenden Schritten (nach Jain u. a. (1999), vergleiche Kapitel 3.2.1):

Festlegung der Datenrepräsentation (*Pattern Representation*) Auswahl einer geeigneten Datenstruktur und vor allem der vom Clustering-Algorithmus zu berücksichtigenden Eigenschaften der Daten. Diese bilden den *Feature Vector*, dessen Elemente (*Features*) quantitative, qualitative oder ordinale Werte oder auch strukturelle Werte wie Bäume beinhalten können. Für diese Arbeit besteht die Struktur der einzelnen Datensätze aus den Metadaten eines Whiskys zur Identifizierung und den zugehörigen, aufbereiteten Tasting Notes.

Definition der Datennähe (*Pattern Proximity*) Um die Nähe beziehungsweise die Distanz zweier Datensätze bestimmen zu können, ist eine geeignete Distanzfunktion zu bestimmen. Diese ist im einfachsten Falle die euklidische Distanz dieser Datensätze:

$$d_2(x_i, x_j) = (\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2)^{1/2}$$

Die Dimension d des euklidischen Raumes entspricht dabei der Anzahl der Möglichen Eigenschaften eines Datensatzes. Um zu verhindern, dass Features mit größeren Skalen andere dominieren, müssen alle Features normalisiert werden.

Clustering Die eigentliche Clustering-Phase dient der Ermittlung der in den Daten beinhalteten Gruppen unter Anwendung der vorher festgelegten Distanzfunktion und Datenrepräsentation. Dabei gibt es verschiedene Vorgehensweisen. Hartes Clustering ordnet jeden Datensatz einer festen Gruppe zu, während unscharfes (*fuzzy*) Clustering zu jedem Datensatz das Maß der Zugehörigkeit zu allen Clustern berechnet. Hierarchisches Clustering baut eine verschachtelte Cluster-Hierarchie auf, während partitionelles Clustering nur eine hierarchische Ebene aufbaut.

Clustering-Methoden unterteilen sich weiter in agglomerative und divisive Verfahren auf. Erstere betrachten jeden Datensatz zunächst als eigenen Cluster und fügt diese zusammen, während letztere einen alle Datensätze umfassenden Cluster immer weiter aufteilen. Monothetische Verfahren vergleichen Datensätze nach und nach anhand einzelner Features, während polythetische die gesamten Feature Vektoren miteinander vergleichen.

Datenabstraktion (*Data Abstraction*) (optional) Die Datenabstraktion dient der besseren Verwertbarkeit der Clustering-Ergebnisse. Die Vorgehensweise richtet sich hierbei nach der weiteren Datenverwendung. Diese kann entweder durch Menschen oder durch weitere automatische Analysen geschehen. Eine typische Abstraktion ist die Generierung einer Clusterbeschreibung. Dies kann beispielsweise ein repräsentativer Datensatz sein.

Bewertung der Ergebnisse (*Assessment of Output*) (optional) Um die Aussagekraft der ermittelten Cluster sicherzustellen, müssen diese im letzten Schritt überprüft werden. Auch hier gibt es verschiedene Vorgehensweisen. Ein relativ sicherer Indikator in der Domäne Whisky sind die bereits bestehenden Kategorien. Sollten die ermittelten Cluster mit diesen Kategorien übereinstimmen, ist davon auszugehen, dass das Ergebnis valide ist. Ein Nachteil dieser Vorgehensweise ist allerdings der Mangel an möglichen neuen Erkenntnissen durch das Messen an bereits bestehendem Wissen. Dieser Nachteil ließe sich durch eine Bewertung des finalen Empfehlungssystems durch möglichst viele Experten verhindern, was allerdings wiederum den Nachteil der Subjektivität dieser Bewertungen mit sich bringt.

3.3 Vektorrepräsentationen von Wörtern

Um die Distanzen zwischen zwei Tasting Notes ermitteln zu können, müssen zunächst die Distanzen der in ihnen enthaltenen Aromen bekannt sein. Diese lassen sich unter anderem entweder über eine Ontologie festlegen oder über einen Algorithmus wie *Word2vec* (Mikolov u. a. (2013b)), welcher eine Erweiterung der *Skip-gram-* und Continuous Bag-of-Words (CBOW) Algorithmen darstellt (Mikolov u. a. (2013a)). Der Vorteil bei letzterer Vorgehensweise liegt in der vergleichsweise weniger aufwändigen Implementierung. Allerdings setzt diese ein ausreichend großes Trainingsset voraus. Dieses Kapitel beschreibt die Grundlagen hinter dem Word2vec-Algorithmus. Die hier getätigten Aussagen basieren vorwiegend auf den beiden genannten Quellen.

Der Ansatz der genannten Algorithmen ist es, Wörter anhand ihrer Kontexte zu vergleichen. Dem liegt die Annahme zugrunde, dass Wörter, welche häufig in einem gleichen oder ähnlichen Kontext verwendet werden, auch eine ähnliche Bedeutung haben (Goldberg und Levy, 2014, S. 5). Je häufiger Wörter in einem ähnlichen Kontext vorkommen, desto wahrscheinlicher ist es, dass sie eine ähnliche Bedeutung haben.

In der Lernphase ermittelt der Algorithmus die Kontexte zu jedem Wort, wobei Kontext die umliegenden Wörter in einem Satz beschreibt. Die Größe des Kontexts ist vom Anwender setzbar. Ein Beispiel für die Kontextbildung bietet Tensorflow (2017). Das so entstehende Set aus Wörtern und deren Kontexten dient als Grundlage für eine Distanzermittlung zwischen den Wörtern. Aus diesem Set bildet der Algorithmus eine Vektorrepräsentation zu jedem Wort, sodass ein multidimensionaler Raum entsteht, in dem ähnliche Wörter, beziehungsweise deren Vektoren, nahe beieinander liegen. Die Dimensionalität dieses Raumes ist ebenfalls variabel setzbar. Ein Beispiel für einen solchen resultierenden Raum ist in Abbildung 3.2 dargestellt.

Eine Erweiterung der ursprünglichen Algorithmen durch wor2vec besteht in der Berücksichtigung von Phrasen. Neben der Kontextgröße und der Dimensionalität sind ebenso Werte für die Mindest- und Höchstanzahl an Vorkommnissen von zu berücksichtigenden Wörtern setzbar, da zu selten oder zu häufig vorkommende Wörter wenig Informationsgehalt haben.

Ein spezielles Resultat dieses Algorithmus ist die Möglichkeit, auf den resultierenden Vektoren logische Operationen durchzuführen. Ein Beispiel ist, dass das Ergebnis der Addition der Vektoren zu „Germany“ und „Capital“ nahe bei dem Vektor zu „Berlin“ liegt. Zu berücksichtigen ist bei diesem Algorithmus, dass die Trainingsdaten das Ergebnis erheblich beeinflussen. So ist es besonders bei wenigen verfügbaren Daten denkbar von Vorteil, ein Trainingsset aus der entsprechenden Domäne zu verwenden.

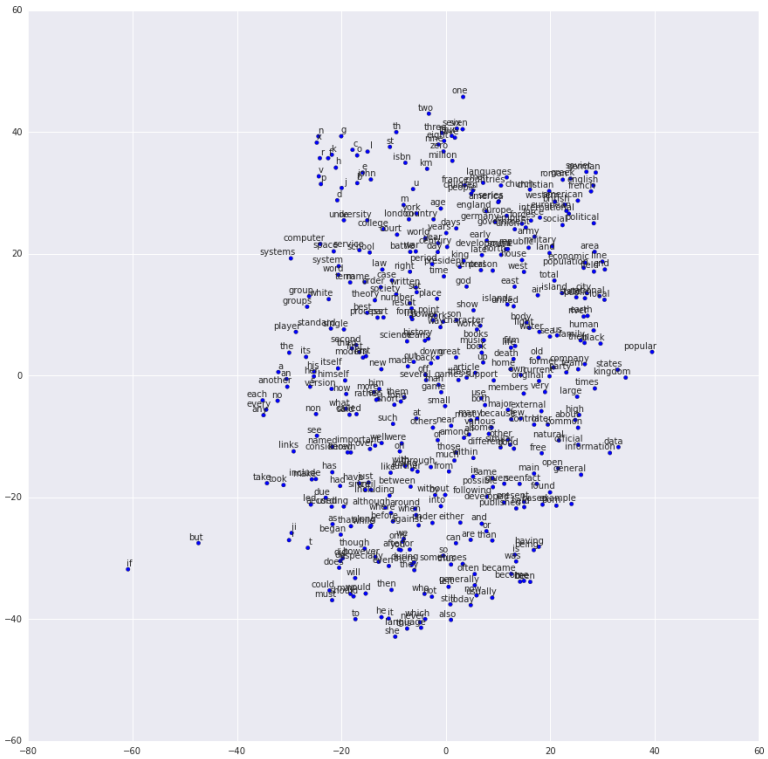

Abbildung 3.2: Beispiel für eine Verteilung von Wörtern im Vektorraum nach einer Dimensionsreduktion auf zwei mit *t-SNE* Tensorflow (2017); Maaten und Hinton (2008)

Eine Umsetzung des Word2vec-Algorithmus in befindet sich in dem *Deeplearning4j*-Plugin (Deeplearning4j Development Team (2017)) für *KNIME* (Berthold u. a. (2007)). Es existiert eine Umsetzung eines Whisky-Empfehlungssystems mit dem Word2vec-Algorithmus (Krzus (2017)). Dieses vergleicht Whisky-Reviews anhand der Cosinus-Distanz der Mittelwerte der Wortvektoren ihrer Tasting Notes zueinander.

4 Aufbau eines Datenkorpus

Dieses Kapitel beschreibt den Aufbau eines Datenkorpus zur Vorbereitung weiterer Experimente. Zunächst müssen nach der Vorgehensweise des KDD-Prozesses geeignete Quellen gesichtet und die ausgewählten Daten bezogen werden. Danach können weitere Schritte erfolgen.

4.1 Vergleich verschiedener Quellen

Bei der Auswahl geeigneter Datenquellen stellen sich mehrere Probleme. Es existieren einige Bücher mit Tasting Notes anerkannter Experten. Diese beinhalten in der Regel jedoch nur wenige hundert Datensätze. Eine Ausnahme bildet die jährlich erscheinende „Whisky Bible“ (Murray (2016)) mit über 4.600 Tasting Notes in der aktuellen Ausgabe. Ein weiteres Beispiel ist das ebenfalls jährlich erscheinende „Malt Whisky Yearbook“ (Ronde (2016)) mit über 200 Tasting Notes. Des Weiteren bietet das Buch eine Liste von Websites, welche teilweise sehr hilfreich sind und ebenfalls Tasting Notes anbieten.

Den Büchern gegenüber stehen Onlinequellen, welche in der Regel Tasting Notes von Amateuren enthalten. Einige Websites werden allerdings von Experten betrieben oder enthalten Tasting Notes von solchen. Scotchwhisky.com (2016) ist ein Onlinemagazin, auf welchem wöchentlich Artikel mit neuen, ausführlichen Tasting Notes erscheinen. Insgesamt sind dort allerdings nur etwas über 500 Tasting Notes vorhanden. Bei den wöchentlich erscheinenden Artikeln werden in der Regel nur speziellere Abfüllungen betrachtet. Die umsatzstärksten beziehungsweise bekanntesten Abfüllungen sind auf der Seite derzeit nicht vorhanden. WhiskyMagazine (2016) ist die Online-Präsenz eines Printmagazins, auf der über 3.000 Whiskys mit Tasting Notes von in der Regel zwei Experten verfügbar sind. WhiskyMonitor (2017) ist eine ausführliche Online-Datenbank mit Daten zu über 16.000 Abfüllungen. Zu über 1.300 davon existieren Tasting Notes.

Für die weitere Arbeit finden zunächst die Daten von WhiskyMonitor (2017) und WhiskyMagazine (2016) Verwendung. Erstere bieten sich besonders dazu an, allgemeine Metadaten zu beziehen. Letztere bieten durch die Tasting Notes bekannter Experten eine geeignete und weniger aufwändig zu beschaffende Alternative zu der „Whisky Bible“.

4.2 Bezug und Pflege der Daten

Für den Aufbau des Datenkorpus und auch für die weitere Arbeit soll eine virtuelle Maschine mit einem entsprechenden virtuellen Environment, welches mit Anaconda (2017) eingerichtet wurde, dienen. Python bietet mit einigen entsprechenden Bibliotheken beziehungsweise Anbindungen an Selenium (2017) und BeautifulSoup (Richardson (2017)) alle nötigen Grundlagen für die Datenbeschaffung.

Der erste Schritt besteht daraus, die Daten von den genannten Seiten zu beziehen und die entsprechenden Metadaten und gegebenenfalls Tasting Notes daraus zu extrahieren. Die so ermittelten Daten liegen zunächst in einer CouchDB-Instanz (CouchDB (2017)). Aus Gründen der Übersichtlichkeit und zum Aufbau eines Metadaten-Konstrukts steht an dieser Stelle die Entscheidung, die Daten in eine MySQL-Datenbank zu übertragen. Zudem erleichtert dies die Anbindung an KNIME.

Eine Möglichkeit, die Qualität der Daten zu erhöhen, liegt darin, die Tasting Notes aus verschiedenen Quellen einander zuzuordnen und somit mehr Daten pro Abfüllung zu haben. Dies ist Ziel eines ersten Versuchs, in dem die Daten aus WhiskyMagazine (2016) und WhiskyMonitor (2017) vereint werden sollen. Um zwei Abfüllungen einander zuordnen zu können, müssen die Metadaten für Destillerie, Abfüller, Marke, Name der Abfüllung, Alter/Reifungsdauer, Herkunft, Ressource und Produktionsweise übereinstimmen. Um Unterschiede in der Schreibweise und Rechtschreibfehler zu umgehen, erweist sich die Methode, bei einigen der Metadaten *slugs* zu verwenden, als hilfreich. Allerdings bereinigt dies nur einen Bruchteil der Unterschiede und Fehler in den Daten. Ein größeres Problem ergibt sich beispielsweise daraus, dass die Daten von WhiskyMagazine (2016) nicht eindeutig zwischen Destillerien und unabhängigen Abfüllern unterscheiden, während die Daten aus WhiskyMonitor (2017) keinen Wert für den Namen der Abfüllung vorsehen, dieser jedoch im Namensfeld der Flasche vorhanden sein kann. Weiter enthielt das ursprüngliche Feld für den Whiskytyp oft Werte wie „japanese single malt“, worin die Werte „Japan“ für die Herkunft, „Single“ für die Produktion und „Malt“ für den Rohstoff enthalten sind. Die Aufsplittung dieser Werte erhöht zwar die Dimensionalität der Metadaten, verringert jedoch die Anzahl der unterschiedlichen Werte in den jeweiligen Feldern erheblich und bietet letztendlich die Möglichkeit, Whiskys nach allen drei Werten zu kategorisieren.

Die beschriebenen Probleme stimmen mit denen überein, welche (Sharafi, 2013, S. 63) nennt. Die Schema-Designs der beiden Seiten stimmen erwartungsgemäß nicht überein, wodurch einiges an Korrekturen und Ergänzungen nötig ist. Je vollständiger die Daten korrigiert sind, desto

genauer und sicherer lassen sich die Datensätze aus den beiden Quellen einander zuordnen und somit die Tasting Notes der Flaschen vereinen.

Da dieser Versuch sehr viel Zeit in Anspruch nimmt, steht an dieser Stelle zunächst die Entscheidung, die Vereinigung der Datensets nicht weiter zu verfolgen und stattdessen mit den Daten aus WhiskyMagazine (2016) als Produktivdatenset und den weiteren Quellen als Trainingsdatenset zu arbeiten. Die bisher ermittelten und bereinigten Metadaten können dennoch im weiteren Verlauf hilfreich sein. Speziell die aufgeteilten Werte für die Whiskytypen lassen sich für eine Validierung der gegebenenfalls ermittelten Cluster verwenden.

5 Zusammenfassung und Ausblick

Diese Arbeit greift das Problem der Whisky-Empfehlungen auf. Es ist unwahrscheinlich, dass ein Mensch auf Anfrage eine akkurate Empfehlung gibt, welche alle möglichen Whiskys berücksichtigt. Für ein besseres Verständnis der Domäne gibt diese Arbeit zunächst eine Einführung in das Thema Whisky. Whisky ist ein breites und vielfältiges Feld mit komplexen aromatischen Zusammensetzungen. Die vielen Aromen entstehen in den verschiedenen Schritten des Fertigungsprozesses. Nicht alle Aromen können auf einen bestimmten Produktionsschritt zurückgeführt werden, weshalb es bis heute bei der Bewertung von Whiskys üblich ist, diese von Menschen durchführen zu lassen. Daher sind diese Bewertungen immer auch zu einem gewissen Grad subjektiv. Bei der Bewertung und Beschreibung ist die Verwendung von Nosing Wheels üblich, welche ein hierarchisches Vokabular bieten. Diese Nosing Wheels bieten zudem eine mögliche Grundlage für eine Aromen-Ontologie.

Empfehlungssysteme sind Software-Tools zur Ermittlung interessanter und nützlicher Items. Sie sind für Nutzer gleichermaßen von Interesse wie für den Anbieter (in der Regel Händler). Es gibt verschiedene Ansätze, Empfehlungssysteme umzusetzen, wobei hier vor allem der Inhaltsorientierte Ansatz von Interesse ist, welcher Items aufgrund ihrer Ähnlichkeit empfiehlt. KDD ist eine allgemeine Prozessbeschreibung der Ermittlung von Wissen aus Rohdaten. Der Prozess beschreibt den Weg vom Verständnis der Domäne über das Data Mining hin zu neu ermitteltem Wissen. Es existieren verschiedene Vorgehensweisen hierbei. Besonders von Interesse ist hier das Clustering als Data-Mining-Methode, welches die automatische Ermittlung von Klassen in Daten beschreibt. Das Clustering teilt sich wiederum in verschiedene Methoden auf. Als ein möglicher Algorithmus für die Generierung der Datenrepräsentation bietet sich Word2vec an.

Weiter beschreibt diese Arbeit die Beschaffung und die Erlangung eines tieferen Verständnisses der verfügbaren Rohdaten und somit die Durchführung der ersten Schritte im KDD-Prozess. Der so entstandene Korpus bietet die Grundlage für weitere Experimente.

In Folge dieser Arbeit soll zunächst ein Teil des aktuellen Datenkorpus dazu dienen, mit dem Word2vec-Algorithmus die Distanzen der in den Tasting Notes verwendeten Geschmacksrichtungen zu berechnen. Als Qualitätsmerkmal für das Ergebnis kann unter anderem das

Nosing-Wheel dienen. Zur Verbesserung der Ergebnisse besteht die Möglichkeit, das Trainingsset zu erweitern und diverse Bereinigungsmethoden anzuwenden. Im Erfolgsfall kann bereits der Versuch beginnen, die Feature-Vektoren der Whiskys miteinander zu vergleichen. Daraufhin muss ein Clustering auf den resultierenden Vektoren durchgeführt werden, um in den Daten liegende Kategorien zu ermitteln. Eine ausführlichere Planung ist in Schole (2017) beschrieben.

Literaturverzeichnis

[Anaconda 2017] Anaconda: *Anaconda*. https://www.continuum.io. 2017. – Letzter Zugriff am 05.03.2017

[Berthold u. a. 2007] Berthold, Michael R. ; Cebron, Nicolas ; Dill, Fabian ; Gabriel, Thomas R. ; Kötter, Tobias ; Meinl, Thorsten ; Ohl, Peter ; Sieb, Christoph ; Thiel, Kilian ; Wiswedel, Bernd: KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer, 2007. – ISBN 978-3-540-78239-1

[CouchDB 2017] CouchDB, Apache: *CouchDB*. http://couchdb.apache.org/. 2017. – Letzter Zugriff am 01.04.2017

[Deeplearning4j Development Team 2017] Deeplearning4j Development Team: *Deeplearning4j: Open-source distributed deep learning for the JVM*. http://deeplearning4j.org. 2017. – Letzter Zugriff am 05.03.2017

[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), November, Nr. 11, S. 27–34. – URL http://doi.acm.org/10.1145/240455.240464. – ISSN 0001-0782

[Goldberg und Levy 2014] Goldberg, Yoav ; Levy, Omer: word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. In: *CoRR* abs/1402.3722 (2014). – URL http://arxiv.org/abs/1402.3722

[Jack 2014] Jack, Frances: Sensory analysis. In: Russell, Inge (Hrsg.) ; Stewart, Graham (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

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

[Krzus 2017] Krzus, Matt: *Whiskey Embeddings*. http://wrec.herokuapp.com/methodology. 2017. – Letzter Zugriff am 05.03.2017

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

[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

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

[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

[Richardson 2017] Richardson, Leonard: *BeautifulSoup*. https://www.crummy.com/software/BeautifulSoup/. 2017. – Letzter Zugriff am 19.02.2017

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

[Schole 2017] 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

[Scotchwhisky.com 2016] Scotchwhisky.com: *Scotchwhisky.com*. https://scotchwhisky.com. 2016. – Letzter Zugriff am 05.03.2017

[Selenium 2017] Selenium: *Selenium*. http://www.seleniumhq.org/. 2017. – Letzter Zugriff am 19.02.2017

[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

[Tensorflow 2017] Tensorflow: *Vector Representations of Words*. https://www.tensorflow.org/versions/master/tutorials/word2vec/. 2017. – Letzter Zugriff am 05.03.2017

[WhiskyMagazine 2016] WhiskyMagazine: *Whisky Magazine*. https://www.whiskymag.com. 2016. – Letzter Zugriff am 05.03.2017

[WhiskyMonitor 2017] WhiskyMonitor: *Whisky Monitor*. https://www.whisky-monitor.com. 2017. – Letzter Zugriff am 05.03.2017

[Wikipedia 2016] Wikipedia: *Adstringenz*. https://de.wikipedia.org/wiki/Adstringenz. 2016. – Letzter Zugriff am 19.10.2016