Whisky Empfehlungen

Joachim Schole

Hochschule für angewandte Wissenschaften, Hamburg, Deutschland
joachim.schole@haw-hamburg.de

Zusammenfassung. Diese Arbeit benennt das allgemeine Problem der Empfehlung von Whiskies basierend auf einem genannten Whisky und befasst sich mit den theoretischen Grundlagen zu Empfehlugnssystemen und Knowledge Discovery in Databases, mit deren Hilfe im weiteren Verlauf der Projekte ein inhaltsbasiertes Empfehlungssystem nach dem KDD-Prozess erschaffen werden soll. Hier stellt die Arbeit die Vision eines geeigneten KDD-Prozesses vor. Weiter geht die Arbeit auf die bisher durchgeführten Experimente und Schritte des KDD-Prozess ein und gibt einen Ausblick auf die angedachten kommenden Schritte. Zuletzt werden mögliche Risiken bei der weiteren Durchführung benannt.

1 Das Problem der Whiskyempfehlung

Wird ein Mensch nach einer Whisky-Empfehlung gefragt, stellt ihn dies vor mehrere Probleme. Er muss sich ausreichend mit der Domäne Whisky auskennen, ausreichend viele Whiskies kennen und dieses Wissen möglichst objektiv und umfassend zu einer geeigneten Empfehlung kombinieren. Hinzu kommt das Einordnen des Geschmacks des Fragenden. Diese Fragen zusammen bilden ein großes, von einem Menschen schwer zu bewältigendes Problem. Hier kann ein Empfehlungssystem helfen, eine alle Möglichkeiten berücksichtigende Empfehlung zu geben.

Whisky ist eine komplexe Domäne. Es herrschen große Unterschiede bezüglich des Geschmacks zwischen Marken und einzelnen Abfüllungen. Besonders erkennbar sind diese zwischen den unterschiedlichen Sorten wie beispielsweise Scotch und Bourbon. Diese Arbeit greift dieses Problem auf und stellt den geplanten Aufbau eines KDD-Prozesses dar, dessen Ergebnis es ermöglichen soll, die geschmackliche Distanz zwischen zwei beliebigen Whiskies aus einem zugrundeliegenden Datenkorpus zu berechnen und somit zu jedem Whisky eine Anzahl von ähnlichsten Whiskies zu ermitteln. Mittels Clustering können zudem optional Kategorien aus den errechneten Distanzen ermittelt werden.

2 Theoretische Grundlagen und Einordnung der Arbeit

Dieses Kapitel beleuchtet die Domäne Whisky und die theoretischen Grundlagen zu Empfehlungssystemen, KDD und Vektorrepräsentationen von Wörtern, welche im geplanten weiteren Verlauf der Experimente notwendig sind.

2.1 Whisky als Domäne

Die Domäne Whisky teilt sich in viele Unterkategorien auf, welche sich vor allem nach Produktionsweise, Herkunft und dem dem Destillat zugrundeliegenden Rohstoff richten. Der grundlegende Ansatz dieser Arbeit ist es allerdings, diese Kategorien bei der Ermittlung einer Ähnlichkeit zwischen Whiskies zunächst zu ignorieren und lediglich die Geschmacksnoten, welche üblicherweise in sogenannten *Tasting Notes* festgehalten werden, zu verwenden. Vielmehr dienen die bestehenden Kategorien als mögliches Kriterium bei der Bewertung der Ergebnisse der Experimente. Eine Tasting Note ist eine textliche, meist sehr kurze geschmackliche Einordnung eines Whiskies und besteht oft nur aus einer Aneinanderreihung von Geschmäckern [5].

Abb. 1. Beispiel für ein Nosing-Wheel [3, S. 238]

Der Geschmack eines Whiskies setzt sich aus einer großen Vielfalt an möglichen Aromen zusammen. Dies ist beispielhaft am *Tasting Wheel* in Abbildung

1 erkennbar. Tasting Wheels sind ein Ansatz, ein einheitliches Vokabular für Tasting Notes zu schaffen [3, S. 237]. Aufgrund der Vielzahl an möglichen Quellen für Tasting Notes ist es jedoch unrealistisch, das Vokabular aus dem Beispiel als ausreichend für die Bildung einer vollständigen Geschmacksontologie zu betrachten.

Die Aromen entstehen in den verschiedenen Produktionsschritten des Whiskies. Hersteller möchten sich in der Regel durch einen eigenen Geschmack von der Konkurrenz abgrenzen und somit einen Wiedererkennungswert schaffen [3, S. 229]. Neben den riechbaren Aromen sind auch die Merkmale süß, bitter und sauer und das Mundgefühl von Bedeutung. Die Entwicklung der Aromen wird durch regelmäßige Proben während des Herstellungsprozess überprüft [3, S. 233]. Eine detailliertere Form der Tasting Notes bilden sogenannte *Flavour Profiles*. Dies sind Bewertungen, welche neben der bloßen Nennung der Aromen auch quantitative Werte zur Intensität dieser beinhaltet. Flavour Profiles bilden eine optimale Grundlage für einen Datenkorpus, da die Verarbeitung numerischer Werte weniger aufwändig ist als die von Texten. Leider existiert keine ausreichend große Datenquelle, welche Flavour Profiles enthält. Daher bilden Tasting Notes in Textform die Datengrundlage für die weiteren Experimente.

2.2 Empfehlungssysteme

Dieses Kapitel beleuchtet theoretische Grundlagen zu Empfehlungssystemen. Die Aussagen basieren im wesentlichen auf [11].

Empfehlungssysteme bieten besonders in Onlineshops eine Hilfe bei der Auswahl des passendsten Produktes. Im Normalfall hat ein Kunde nicht die Muße, sich sämtliche in Frage kommenden Produkte eines Händlers anzusehen. Vielmehr ist es meistens der Anspruch, möglichst schnell zum Ziel zu kommen. Zudem helfen Empfehlungssysteme speziell den Menschen, die sich in der Domäne des gesuchten Artikels gar nicht oder wenig auskennen.

Ein Empfehlungssystem betrachtet Objekte von Interesse, *Items* genannt. In dieser Arbeit sind diese Objekte Whiskies beziehungsweise Whiskyabfüllungen. Oft ist ein Empfehlungssystem auf eine spezielle Domäne ausgerichtet, um durch Spezifizierung bessere Ergebnisse zu erzielen. Je nach Domäne und Beschaffenheit der Rohdaten bieten sich verschiedene Methoden der Empfehlungsgenerierung an. Hierbei liegen in der Regel Daten der drei Typen Item, Nutzer und Transaktion zugrunde, wobei eine Transaktion eine Aktion zwischen einem Nutzer und einem Item beschreibt, wie beispielsweise ein Kauf oder eine Bewertung.

Die wesentlichen möglichen Methoden der Empfehlungsgenerierung sind *content-based*, *collaborative filtering*, *demographic*, *knowledge based*, *community based* und *hybrid*, wobei letzteres eine beliebige Kombination der vorherigen meint. Für diese Arbeit von Interesse ist zunächst der Inhaltsbasierte (content based) Ansatz, welcher Items allein anhand ihrer Eigenschaften vergleicht. In weiteren denkbaren Entwicklungen der Arbeit sind möglicherweise das Collaborative Filtering oder der Community Based Ansatz interessant, welche die Vorlieben von Personen mit ähnlichem Geschmack wie der Kunde beziehungsweise Personen aus dessen sozialen Umfeld als Empfehlungsgrundlage verwenden.

2.3 Knowledge Discovery in Databases

Knowledge Discovery in Databases beschreibt einen allgemeinen Prozess zur Wissensgewinnung aus Rohdaten [2,1]. Der Prozess ist in Abbildung 2 dargestellt.

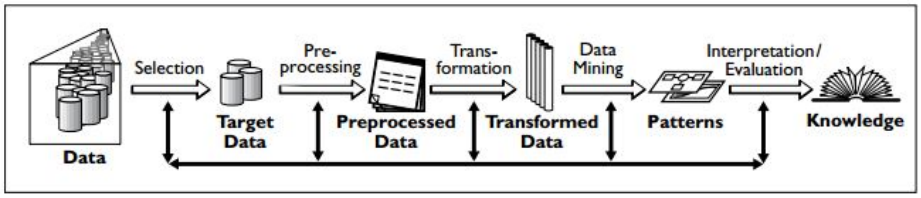

Abb. 2. Der KDD-Prozess [2, S. 29]

Er beginnt mit der Aneignung von ausreichend tiefem Domänenwissen. Dieses ist durch die Befassung mit der Domäne Whisky weitestgehend abgeschlossen. Da dieser Schritt grundlegend für die weitere Arbeit ist und die Benennung des Ziels des KDD-Prozesses und der zugehörigen Risiken erst möglich macht, kann er einen Großteil der gesamten Arbeit einnehmen [2]. Mit dem angeeigneten Domänenwissen kann darauf folgend die Auswahl geeigneter Rohdaten erfolgen. Im bisherigen Projektverlauf wurde sich bereits mit einigen möglichen Datenquellen befasst und eine Auswahl getroffen. Dies ist in Kapitel 3 näher beschrieben. Die Rohdaten müssen im weiteren Verlauf bereinigt und vorverarbeitet werden. Eine solche Bereinigung ist ebenfalls bereits geschehen. Bestandteile dieses Schritts sind das Entfernen unbrauchbarer und fehlerhafter Daten und die Rauscherkennung [14, S. 62 ff.]. Mögliche weitere Probleme können auf Schema- und Instanzlevel auftauchen. Besonders bei mehreren Datenquellen können heterogene Datenmodelle ein Hindernis auf schematischer Ebene sein [14, S. 63]. Genau dieser Umstand bildet die größte Herausforderung bei der Ermittlung von Datensätzen aus verschiedenen Quellen, welche den gleichen Whisky beschreiben.

Im nächsten Schritt erfolgt die Transformation der Daten in eine geeignete Repräsentationsform. Folgend müssen geeignete Data-Mining- Methoden ermittelt und daraufhin durchgeführt werden. Mögliche Typen von Data-Mining-Methoden sind die Klassenbildung (*Clustering*), die Klassifizierung (*Classification*), die Assoziationsanalyse und die Zeitreihenanalyse [14, S. 6 ff.]. Für diese Arbeit sind vor allem Methoden des Typs Clustering von Interesse, welches die automatische Erkennung von in den Daten liegenden Klassen durch Partitionierung beschreibt.

Das Clustering beinhaltet ebenfalls den Schritt der Transformation der Daten in eine Repräsentationsform [4, S. 270 f.]. Folgend muss die Ermittlung der Nähe der Datensätze festgelegt werden [4, S. 267]. Dies ist weitestgehend geschehen und ebenfalls in Kapitel 3 beschrieben. Die Durchführung des eigentlichen

Clusterings führt zur Ermittlung der Klassen. Dabei bestehen wieder mehrere mögliche Methoden.

Die so entdeckten Patterns (Die Distanzen der Whiskies und evtl. die Klassen) in den Daten müssen dann zur Wissensgewinnung interpretiert werden. Auf Grundlage dieses Wissens können Maßnahmen getroffen werden, zu denen auch eine Anpassung des KDD-Prozesses gehören kann.

2.4 Vektorrepräsentationen von Wörtern

Eine naheliegende Möglichkeit zur Ermittlung der Distanzen zwischen Whiskies ist die, zunächst die Distanzen zwischen den möglichen Geschmacksrichtungen zu ermitteln und mittels dieser Distanzen anschließend die Distanzen der Whiskies, welche dann eine Menge von Geschmäckern bilden, zu berechnen.

Ein geeigneter Kandidat hierfür ist der *word2vec*-Algorithmus [8]. Dies ist eine Erweiterung der Skip-Gram und Continuous Bag-of-Words (CBOW) Algorithmen [7]. Im Gegensatz zu anderen üblichen NLP-Systemen verfolgen die Algorithmen den Ansatz, Wörter anhand ihrer Kontexte zu vergleichen und somit eine Ähnlichkeit zu ermitteln [7]. Den Algorithmen liegt die Annahme zugrunde, dass Wörter, die häufig in ähnlichen Kontexten verwendet werden, auch eine ähnliche Bedeutung haben. In der Trainingsphase ermittelt der Algorithmus die Kontexte jedes Wortes im Trainingsset. Der Kontext beschreibt hier die umliegenden Wörter in einem Satz. Die Größe des Kontexts ist setzbar.

So entsteht ein Set aus Wörtern und deren Kontexten, woraus sich ermitteln lässt, wie wahrscheinlich es ist, dass aus einem Kontext auf ein bestimmtes Wort zu schließen ist (CBOW), oder dass aus einem Wort auf den Kontext geschlossen werden kann (Skip-Gram) [7]. Auf Basis der Kontexte bildet der Algorithmus zu jedem Wort eine Vektorrepräsentation. Die Größe des Kontexts hat einen Einfluss auf die Ergebnisse. Ein größerer Kontext erhöht die Genauigkeit der Vorhersagen [7].

Die Erweiterung des Skip-Gram Algorithmus durch word2vec besteht unter anderem in der initialen Suche im Trainingsset nach Phrasen, welche oft eine andere Bedeutung haben, als die einzelnen Wörter [8]. Diese werden in der Folge wie Wörter behandelt. Jedes Wort aus dem Trainingsset wird nun durch seine Kontexte repräsentiert und es kann ermittelt werden, in welchen Kontexten das Wort am häufigsten vorkommt. Der Vergleich der Kontexte zweier Wörter bildet somit ein Ähnlichkeitsmaß dieser Wörter. Als Ergebnis des Trainings steht ein Set von Wörtern und jeweils einer Vektorrepräsentation dieser. Die Dimensionalität der Vektoren ist ebenfalls setzbar. Eine höherer Wert für die Dimensionalität führt auch hier zu besseren Ergebnissen. Wörter, deren Vektoren in diesem Raum nahe beieinander liegen kommen häufig in ähnlichne Kontexten vor und haben somit nach der oben genannten Annahme auch eine ähnliche Bedeutung.

Übertragen auf Tasting Notes ermittelt dieser Algorithmus die Ähnlichkeiten zwischen Geschmacksrichtungen. Auf Grundlage dieser Ähnlichkeiten kann dann die Ähnlichkeit der summierten Geschmacksrichtungen ermittelt werden.

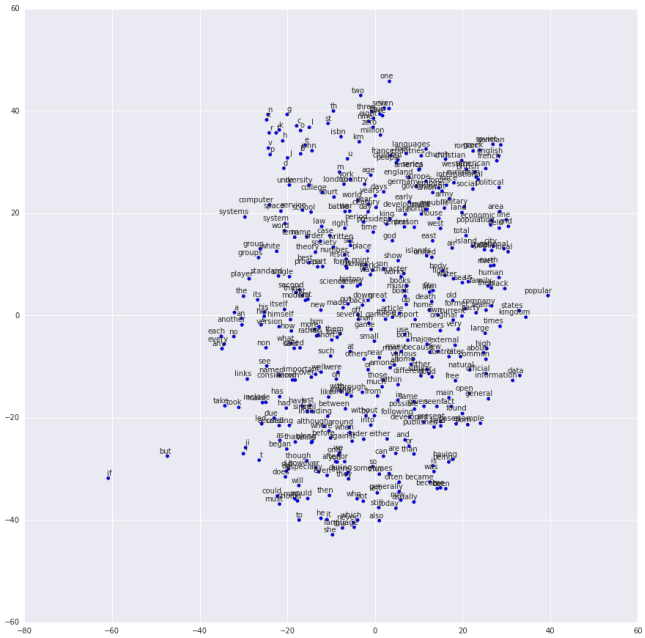

Abb. 3. Beispiel für eine Verteilung von Wörtern im Vektorraum nach einer Dimensionsreduktion auf zwei mit *t-SNE* [16,6]

2.5 Weitere relevante Arbeiten und Abgrenzung

Neben den bisher genannten Arbeiten ist *wrec.herokuapp.com* besonders von Interesse [5]. Hier ist die Funktionsweise eines Whisky-Empfehlungssystems mit word2vec beschrieben. Zudem ist dies die einzige bisher gefundene Arbeit, welche sich mit Whisky auf der Ebene des Text Mining befasst.

Der Unterschied zu dieser Arbeit liegt unter anderem im Datenset. Das hier verwendete Set stammt aus mehreren Quellen und ist bisher noch kleiner, während das genannte System lediglich Tasting Notes von *Reddit.com* verwendet [5,10] Weiter verwendet die Arbeit den Median der Wortvektoren eines Whiskies als Datenrepräsentation. An dieser Stelle verfolgt diese Arbeit die Idee, andere Werte als Datenrepräsentation zu verwenden. So ist es denkbar, den Mittelwert der Wortvektoren eines Whiskies zu verwenden. Weiter bestünde die Möglichkeit, die Wortvektoren bei der Summierung nach der Häufigkeit ihres Vorkommens in verschiedenen Tasting Notes zu gewichten. Dies erscheint allerdings erst ab einer größeren Anzahl von Tasting Notes (ab 5 oder 10) sinnvoll, da bei den hier bisher verwendeten maximalen zwei Tasting Notes pro Whisky sämtliche Geschmacksnoten entweder mit 0.5 oder 1 gewichtet würden. Um dies zu verhindern, kann eine angepasste Gewichtung verwendet werden.

Ein weiterer Unterschied ist das (mögliche) Clustering zur Ermittlung von Kategorien im Datenset. Dies würde eine neue Ebene hinzufügen und es ermöglichen, diese Kategorien mit bestehenden zu vergleichen.

Die Idee dieser Experimente baut auf der Masterthesis von Sigurd Sippel auf, welche sich mit Cocktail-Empfehlungen befasst [15]. Allerdings stellt sich schnell heraus, dass die dort durchgeführten Experimente nicht oder nur teilweise auf die Domäne Whisky übertragbar sind. So ist es sinnvoll, einen KDD-Prozess zu verwenden. Aufgrund der unterschiedlichen Beschaffenheit der Daten ist es allerdings nicht möglich, die dort angewandten Data-Mining-Methoden direkt zu übertragen. Daher erscheint es für die Domäne Whisky sinnvoller, mit Methoden wie den in Kapitel 2.4 genannten Algorithmen zu arbeiten.

3 Aktueller Stand der durchgeführten Experimente

Die bisherige Projektarbeit umfasst die vertiefte Recherche zu den benötigten Technologien und die Einarbeitung in die Domäne Whisky, um einen geeigneten KDD-Prozess skizzierten zu können. Diese Schritte sind in Kapitel 2 skizziert.

3.1 Auswahl der Datenquellen

Nach dem KDD-Prozess erfolgt nach der Einarbeitung in die Domäne die Auswahl der Rohdaten. Dies beinhaltet die Betrachtung und den Vergleich verschiedener möglicher Datenquellen. Neben physischer Literatur wie der *Whisky Bible* [9] und dem *Malt Whisky Yearbook* [12] sind vor allem Online-Quellen wie *whiskymag.com* [17], *whisky-monitor.com* [18] und *scotchwhisky.com* [13] zu betrachten. Es existieren weitere Quellen wie das in Kapitel 2.4 genannte Reddit.com. Für diese Arbeit finden allerdings bisher nur die genannten Quellen betrachtung. Während Bücher Tasting Notes von weitgehend anerkannten Experten beinhalten, bieten Online-Quellen eine deutlich größere Menge an Daten und eine verhältnismäßig weniger aufwändige Beschaffung dieser. Eine Herausforderung bei dem Bezug der Daten aus Büchern ist die in der Regel ausschließlich in physischer und nicht in digitaler Form verfügbare Literatur. Nach dem Vergleich der Quellen steht die Entscheidung, für diese Arbeit auf Online-Quellen zurückzugreifen.

3.2 Beschaffung der Daten

Der nächste Schritt im Prozess sieht die Beschaffung der ausgewählten Daten vor. Die Daten der oben genannten Online-Quellen müssen mittels Webscraping bezogen und weiter betrachtet werden. Hierbei gilt es, die Eignung der Daten als Grundlage für ein Empfehlungssystem zu bewerten. Hauptkriterien sind die Größe des Datensets und der Ursprung der Tasting Notes. In jedem Fall kommen die Daten als Trainingsset in Frage.

Scotchwhisky.com behandelt fast ausschließlich speziellere Abfüllungen und bietet eine Datenmenge von unter 600 Tasting Notes und ist damit als Datengrundlage für ein Empfehlungssystem ungeeignet. *Whisky-monitor.com* bietet

über 3.400 Tasting Notes zu über 2.300 Abfüllungen einer Gruppe von Amateuren und kommt daher als Datenset in Frage. Insgesamt dient die Seite als Whisky-Datenbank mit über 16.000 Abfüllungen, was sie zu einer geeigneten Grundlage für ein Metadatenset macht. *Whiskymag.com* ist die Online-Präsenz eines Printmagazins mit Tasting Notes von in der Regel zwei Experten zu über 3.300 Whiskies. Insgesamt beläuft sich die Zahl der Tasting Notes auf dieser Seite nach der Bereinigung auf über 6.500. Damit ist diese Seite als Quelle für das Empfehlungssystem am besten geeignet.

3.3 Vorverarbeitung der Daten

Nach der Datenauswahl und Beschaffung erfolgt die Vorverarbeitung. Zur Verfeinerung des Geschmacksprofils der Whiskies sollen die Tasting Notes der verschiedenen Quellen zusammengeführt werden. Hierfür muss eine Bereinigung des Korpus um unbrauchbare Daten, eine Korrektur fehlerhafter und gegebenenfalls eine Erweiterung fehlender Metadaten erfolgen. Stimmen die Daten Destillerie, Abfüller, Marke, Alter, Name der Abfüllung, Herkunftsregion, Rohstoff und Produktionsweise zweier Whiskies überein, kann davon ausgegangen werden, dass sie dieselbe Flasche beschreiben. Aufgrund schematischer Unterschiede zwischen den beiden Datensets erweist sich dies als unerwartet schwierig (vgl. [14]). Daher steht an diesem Punkt die Entscheidung, die Datensets zunächst getrennt als Trainings- und Live-Datenset zu verwenden.

Auf den bisherigen Daten wurden erste Versuche, ein Geschmacksvokabular mittels word2vec zu bilden, durchgeführt. Diese Versuche stehen allerdings noch am Anfang und bringen bisher keine brauchbaren Ergebnisse hervor.

Damit steht die Arbeit im Sinne des KDD-Prozesses zwischen dem Schritt der Datenaufbereitung und dem Schritt der Datentransformation, wobei der KDD-Prozess durch seinen iterativen Aufbau immer Rückschritte zu Korrekturzwecken erlaubt [2].

4 Skizze der zukünftigen Arbeiten

Die bisherige Arbeit ist in Kapitel 3 beschrieben. Dieses Kapitel skizziert den geplanten weiteren Verlauf der Experimente im Sinne des KDD-Prozesses.

4.1 Datentransformation

Anschließend an den beschriebenen Aufbau des Datenkorpus erfolgt die Datentransformation. Hierfür muss mit word2vec ein Vokabular geschaffen werden, welches die Ähnlichkeiten der möglichen Geschmacksrichtungen darstellt. Dieses Vokabular kann beispielsweise anhand des Nosing-Wheels qualitativ bewertet werden. Hier gilt es, zu testen, wie sich das Vokabular verbessert beziehungsweise verschlechtert, wenn Parameter wie die Kontextgröße der Wörter beim Training

angepasst werden. Gegebenenfalls muss das Trainingsset durch weitere Online-Quellen vergrößert werden. Auf Grundlage des so entstehenden Geschmacksraumes können die Geschmacksvektoren der Whiskies berechnet werden. Hierbei gilt es, eine geeignete Methode zur Berechnung dieser zu ermitteln.

4.2 Ermittlung einer Distanzfunktion

Der nächste Schritt im KDD-Prozess ist die Ermittlung einer Distanzfunktion. Es existieren mehrere Möglichkeiten, die Distanzen zwischen Vektoren zu berechnen. Hier gilt es, die geeignetste zu ermitteln. In der Regel wird im Umfeld der Vektorrepräsentationen die Cosinus-Distanz verwendet [8,5]. Gemäß der Vision der Arbeit ist es an diesem Punkt bereits möglich, Empfehlungen zu Whiskies zu geben und diese qualitativ zu bewerten. Anhand der Positionen der Geschmacksvektoren im Geschmacksraum kann zusätzlich noch ein Clustering durchgeführt werden. Ein Bewertungskriterium der Ergebnisse wäre, ob die beim Clustering ermittelten Kategorien mit bestehenden Kategorien in der Domäne übereinstimmen.

4.3 Validierung der Ergebnisse

Abschließend muss eine Validierung der ermittelten Distanzen erfolgen. Um die Qualität der Ergebnisse sicher validieren zu können, müssen sie einem Publikum zur Verfügung gestellt werden, welches die Empfehlungen bewerten kann. Hierbei stellt sich die Herausforderung, dass die meisten Menschen nicht ausreichend mit der Domäne Whisky befasst sind. Zudem ist es nicht möglich, eine ausreichend große Gruppe von Experten zu befragen.

Die momentan bevorzugte Idee ist es, eine Webpräsenz aufzubauen, welche die Daten intuitiv und verständlich darstellt und Nutzern erlaubt, diese zu bewerten. Ein einfacher Ansatz hierfür ist eine Suchmaschinen-ähnlich aufgebaute Seite, welche einen Whisky als Eingabe akzeptiert und die ähnlichsten n Whiskies zurückgibt. Die Seite bietet dem Nutzer daraufhin an, jede dieser Empfehlungen einzeln positiv, negativ oder neutral/unbekannt zu bewerten. Eine Herausforderung besteht hier in der Vorbeugung von Missbrauch einer solchen Webpräsenz. Weiterhin müssen die Nutzer der Seite nach ihrer Erfahrung mit Whisky in verschiedene Kategorien aufgeteilt werden, um ihre Bewertungen besser einordnen zu können. Hierfür ist es notwendig, die Nutzer einen Account anlegen zu lassen. Gegebenenfalls können so die Bewertungen eines Nutzers komplett entfernt werden.

Zu den weiteren optionalen Schritten gehört die Weiterverfolgung des Versuchs, mehrere Quellen zu vereinen und so mehr Tasting Notes pro Whisky verarbeiten zu können. Ein wahrscheinlich notwendiger Schritt ist die Vergrößerung des Trainingssets für die Erstellung der Wortvektoren.

5 Bestehende Risiken bei den verbleibenden Experimenten

Neben der Zeit im Allgemeinen betrifft ein weiteres Risiko den geplanten Schritt der Vokabularbildung. Hierbei kann sich herausstellen, dass die verwendeten Daten in der vorhandenen Menge nicht ausreichend sind und neue Quellen zusätzlich bezogen werden müssen. Sollte sich herausstellen, dass diese Methode zu gänzlich ungeeigneten Ergebnissen führt, bestünde die Alternative, auf Grundlage des Nosing-Wheels eine hierarchische Geschmacksstruktur aufzubauen. Hierbei besteht allerdings das Problem, dass sämtliche in den Quellen vorkommende Geschmacksrichtungen berücksichtigt werden müssen und das Nosing-Wheel vermutlich manuell erweitert werden müsste.

Weiterhin besteht das Risiko, dass die angestrebte Validierungsmethode aus zeitlichen oder anderen Gründen nicht umgesetzt werden kann. In diesem Fall muss alternativ gezielt eine Anzahl an freiwilligen Personen mit ausreichender Fachkenntnis ermittelt werden, welche die Ergebnisse validieren.

Die in Kapitel 4 genannten Schritte sind für den weiteren Verlauf der Projekte geplant. Abhängig vom Fortschritt kann an unterschiedlichen Punkten nach der abgeschlossenen Berechnung der Distanzen mit der Masterthesis begonnen werden. Es wird angestrebt, dass die Validierungsphase der Ergebnisse möglichst mit der Aufnahme der Arbeit an der Thesis beginnt und somit während dieser lediglich die Auswertung der Bewertungen als praktischer Teil bleibt.

6 Zusammenfassung

Diese Arbeit befasst sich mit dem allgemeinen Problem der Empfehlung eines Whiskys. Zur Lösung des Problems stellt die Arbeit den geplanten Ansatz eines KDD-Prozesses zur Ermittlung der Ähnlichkeiten beziehungsweise Distanzen zwischen Whiskies vor. Dieser KDD-Prozess soll die Grundlage für ein Empfehlungssystem bilden. Kern des KDD-Prozesses bildet die Transformation der Textdaten in Vektorform unter Verwendung des word2vec-Algorithmus und der Vergleich dieser. Als Datenkorpus dienen Daten aus verschiedenen Online-Quellen. Die bisherigen Arbeiten umfassen die eingehende Beschäftigung mit der Domäne Whisky und die Betrachtung verschiedener in Frage kommender Datenquellen. Weiterhin die Beschaffung der Daten aus den ausgewählten Quellen, die Bereinigung dieser und erste experimentelle Analysen.

Im weiteren Verlauf der Experimente sollen diese Analysen konkretisiert werden. Explizit bedeutet dies eine verfeinerte Abstimmung der Algorithmen auf das Datenset und eine bessere Bereinigung der Daten. Die bei den Experimenten entstandenen Ergebnisse gilt es zu validieren. Die Risiken bei den kommenden Experimenten betreffen vor allem die Zeit und die Eignung der gewählten Algorithmen. Gegebenenfalls müssen Korrekturen am KDD-Prozess vorgenommen werden.

Literatur

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996), http://dx.doi.org/10.1609/aimag.v17i3.1230
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (Nov 1996), http://doi.acm.org/10.1145/240455.240464
3. Jack, F.: Sensory analysis. In: Russell, I., Stewart, G. (eds.) Whisky, pp. 229 – 242. Academic Press, San Diego, second edition edn. (2014), http://www.sciencedirect.com/science/article/pii/B9780124017351000131
4. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Comput. Surv. 31(3), 264–323 (Sep 1999), http://doi.acm.org/10.1145/331499.331504
5. Krzus, M.: Whiskey embeddings. http://wrec.herokuapp.com/methodology (2017), letzter Zugriff am 05.03.2017
6. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality
9. Murray, J.: Jim Murray's Whisky Bible 2017. Dram Good Books (2016)
10. Reddit: Reddit. https://www.reddit.com/ (2017), letzter Zugriff am 05.03.2017
11. Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook, pp. 1–35. Springer US, Boston, MA (2011), http://dx.doi.org/10.1007/978-0-387-85820-3_1
12. Ronde, I.: Malt Whisky Yearbook 2017. MapDig Media Limited (2016)
13. Scotchwhisky.com: Scotchwhisky.com. https://scotchwhisky.com (2016), letzter Zugriff am 05.03.2017
14. Sharafi, A.: Knowledge Discovery in Databases, pp. 51–108. Springer Fachmedien Wiesbaden, Wiesbaden (2013), http://dx.doi.org/10.1007/978-3-658-02002-6_3
15. Sippel, S.: Domain-specific recommendation based on deep understanding of text. Ph.D. thesis, Hochschule für angewandte Wissenschaften Hamburg (April 2016)
16. Tensorflow: Vector representations of words. https://www.tensorflow.org/versions/master/tutorials/word2vec/ (2017), letzter Zugriff am 05.03.2017
17. WhiskyMagazine: Whisky magazine. https://www.whiskymag.com (2016), letzter Zugriff am 05.03.2017
18. WhiskyMonitor: Whisky monitor. https://www.whisky-monitor.com (2017), letzter Zugriff am 05.03.2017