Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Hauptprojekt

Joachim Schole

Whisky-Empfehlungen

Evaluierung eines KDD-Prozesses zur Entwicklung eines Empfehlungssystems

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Joachim Schole

Whisky-Empfehlungen

Evaluierung eines KDD-Prozesses zur Entwicklung eines Empfehlungssystems

Ausarbeitung Hauptprojekt eingereicht im Rahmen des Hauptprojekt

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: 4. Februar 2018

Inhaltsverzeichnis

1. Einleitung

Eine Empfehlungsanfrage bildet immer eine komplexe Problemstellung. Der Gefragte muss über ausreichende Domänenkenntnisse verfügen und den Geschmack des Fragenden in dieses Wissen einordnen können. Die Domäne Whisky bildet hier keine Ausnahme. Ziel dieser Arbeit ist es daher, die Grundlagen eines Whisky-Empfehlungssystems zu schaffen. Konkret beschreibt diese Arbeit einen ersten Durchlauf eines in vorigen Arbeiten geplanten und vorbereiteten KDD-Prozesses Schole (2017a),Schole (2017b).

Die genannten Arbeiten beschreiben die theoretischen Grundlagen der Arbeit und die Herleitung des KDD-Prozesses. Dieser ist in Kapitel 2 näher beschrieben.

Empfehlungssysteme sind Systeme, welche Anwendern anhand verschiedener Kriterien Objekte von Interesse vorschlagen. In diesem Fall soll ein rein Inhaltsbasiertes Empfehlungssystem entwickelt werden. Das bedeutet, es werden lediglich die Eigenschaften der untersuchten Objekte miteinander verglichen und keine übergeordneten Daten wie beispielsweise Bewertungen von Nutzern. Das Ziel ist es, Whiskys allein anhand der zu ihnen verfassten Tasting Notes zu vergleichen.

2. Ausgangslage

In den bisherigen Arbeiten ist ein auf die Problemstellung zugeschnittener KDD-Prozess entwickelt worden und in ersten praktischen Versuchen ein vorläufiger Datenkorpus entstanden. Im folgenden ist der geplante KDD-Prozess beschrieben.

2.1. Aufbau des KDD-Prozesses

Der KDD-Prozess für diese Arbeit orientiert sich am allgemeinen KDD-Prozess aus Fayyad u. a. (1996b). Abbildung 2.1 zeigt den grundlegenden Aufbau eines KDD-Prozesses. Im folgenden sind die einzelnen Schritte mit konkretem Bezug zu diesem Projekt näher erläutert. Eine allgemeine Beschreibung der Schritte bieten die Vorigen Arbeiten Schole (2017a) und Schole (2017b).

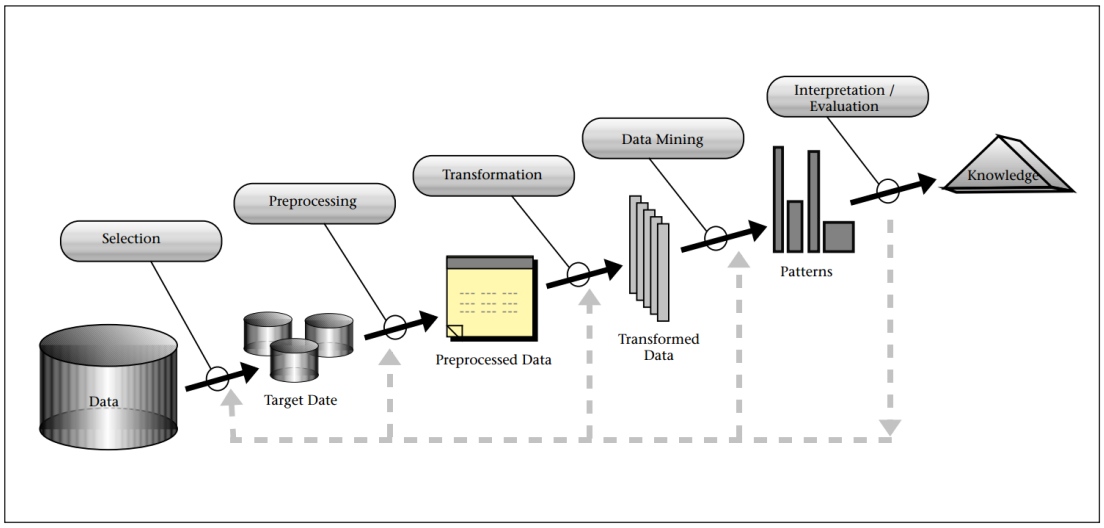

Abbildung 2.1.: Der KDD-Prozess (Fayyad u. a., 1996a, S. 41)

Datenselektion Die Datenselektion ist durch eine Sichtung verschiedener physischer und digitaler (Online-) Quellen erfolgt, an deren Ende die Entscheidung für letztere steht. Die

ausgewählten Online-Quellen sind daraufhin per Web Scraping in eine lokale Datenbank geladen und strukturell aufbereitet worden. Dabei ist noch keine Vorverarbeitung der Texte an sich erfolgt.

Vorverarbeitung In einem ersten Ansatz sollen Tasting Notes aus verschiedenen Quellen, welche denselben Whisky beschreiben, einander zugeordnet werden. Dies erweist sich als sehr aufwändig, weshalb die Entscheidung fällt, dass die Tasting Notes aus der Quelle WhiskyMagazine (2016) als Testdatenset dient und die Tasting Notes aus allen anderen Quellen als Trainingsdatenset dienen (WhiskyMonitor (2017); Scotchwhisky.com (2016); WhiskyIntelligence (2017)).

Die weitere Vorverarbeitung der Daten beinhaltet die gängigen Text-Mining-Methoden wie das Stopword-Removal und das Stemming. Als zentraler Algorithmus für die Vorverarbeitung der Trainingsdaten soll der Word2Vec-Algorithmus Mikolov u. a. (2013) dienen. Dieser ermittelt aus einem Trainingsdatenset die Wahrscheinlichkeiten, mit denen von einem Wort auf den restlichen Kontext in einem Satz geschlossen werden kann und umgekehrt. Die so entstehenden Vektorrepräsentationen der Wörter bilden einen multidimensionalen Raum, in dem synonyme Wörter nahe beieinander liegen. Der gedankliche Ansatz hinter diesem Algorithmus ist die Annahme, dass Wörter, die häufig in ähnlichen Kontexten vorkommen auch ähnliche Bedeutungen haben. Der Word2Vec-Algorithmus arbeitet mit Sätzen, weshalb die Trainingsdaten zusätzlich noch in Sätze aufgeteilt werden müssen. Die so entstehenden Wortvektoren bilden die Grundlage für die anschließende Transformation.

Für die Transformation müssen die Daten aus dem Testdatenset ebenfalls mindestens per Stemming vorverarbeitet werden, damit beide Sets auf Wortstämme reduziert sind und eine korrekte Zuordnung der Wortvektoren möglich ist. Das Stopword-Removal ist hier theoretisch nicht von Bedeutung, da lediglich die Wörter, zu denen ein Vektor existiert, in die Repräsentationsform mit einfließen.

Transformation Die Transformation der Daten erfolgt durch die Anwendung der Wortvektoren auf die Tasting Notes aus dem Testdatenset. Zu jedem Whisky werden die ihn beschreibenden Wortvektoren auf Grundlage der zugehörigen Tasting Note ermittelt und aus diesen der Median gebildet. Der so errechnete Vektor bildet die Repräsentationsform des Whiskys. Die Distanz beziehungsweise Nähe dieser Vektoren stellt die geschmackliche Distanz der zugehörigen Whiskys dar.

Data Mining Als Data-Mining-Methode dient das Clustering auf Basis der Distanzen der Whisky-Vektoren zueinander. Hierfür müssen diese Distanzen zunächst berechnet werden, woraufhin das Clustering durchgeführt werden kann. Als erster Ansatz für diesen Schritt dient hier die Berechnung einer Distanzmatrix auf Basis der Cosinus-Distanzen der Whisky-Vektoren und ein darauf basierendes hierarchisches Clustering. Alternativ bietet sich als Distanzmaß auch die euklidische Distanz an.

Interpretation Zuletzt müssen die so ermittelten Cluster überprüft werden und mögliche Optimierungsmaßnahmen geplant werden.

3. Erster Durchlauf des KDD-Prozesses

Zur Durchführung der Experimente dient die Software KNIME (Berthold u. a. (2007)). In dieser ist bereits eine Implementierung des Word2Vec-Algorithmus enthalten (Deeplearning4j Development Team (2017)). KNIME bietet die Möglichkeit, verschiedene Datenverarbeitungsalgorithmen relativ einfach zu einem *Workflow* zusammenzuschließen. Zunächst soll ein grober Workflow aufgebaut werden, welcher dem zuvor beschrieben KDD-Prozess folgt. Der Workflow teilt sich aus Performanz- und Sicherheitsgründen in mehrere Unterschritte auf.

3.1. Training der Wortvektoren

Das Training der Wortvektoren geschieht durch den Word Vector Learner Node in KNIME, welcher eine Implementierung des Word2Vec-Algorithmus darstellt. Die Ergebnisse dieses Nodes lassen sich mit dem Word Vector Model Extractor Node auslesen und betrachten. Weiterhin lassen sich die Ergebnisse mittels t-SNE visualisieren und optisch evaluieren. Für letzteren Schritt findet eine Implementierung des t-SNE-Algorithmus in Python Verwendung, welche über einen Python Snippet Node eingebunden wird (Ter Heide (2017); van der Maaten (2017)). Abbildung 3.1 zeigt den KNIME-Workflow für das Wortvektoren-Training.

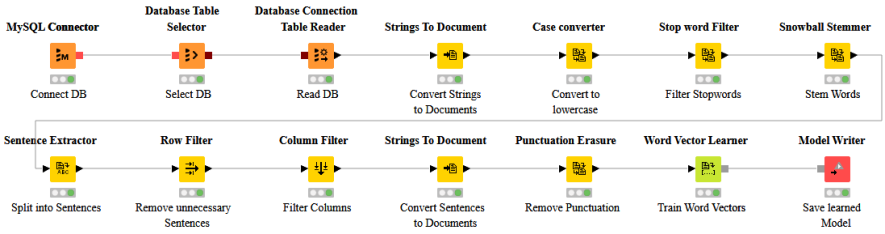

Abbildung 3.1.: KNIME-Workflow für das Wortvektoren-Training

Abbildung 3.2 zeigt das Ergebnis der Anwendung des Word Vector Learners auf die Rohdaten. Abbildung 3.3 zeigt das Ergebnis nach vorheriger Aufteilung der Texte in Sätze, wie es der Anwendung des Word2Vec-Algorithmus entspricht. Abbildung 3.4 zeigt das Ergebnis nach vorheriger Anwendung des Standard-Stopwordfilter-Nodes aus KNIME und des Snowball-Stemmer-Nodes mit dem Kuhlen-Stemmer. Es ist eine klare Verbesserung zu den vorigen

Grafiken erkennbar. Es lässt sich bereits der Trend erkennen, dass Wörter, welche häufig zur Beschreibung von Whiskys verwendet werden, sich in der Linienförmigen Ansammlung oben wiederfinden.

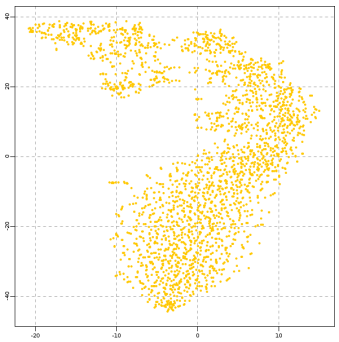

Abbildung 3.2.: Rohdaten

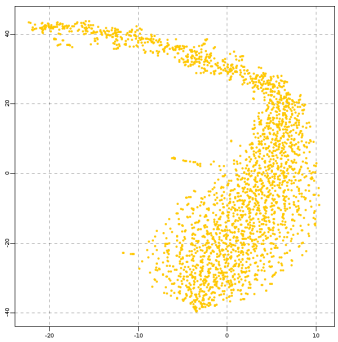

Abbildung 3.3.: Sätze

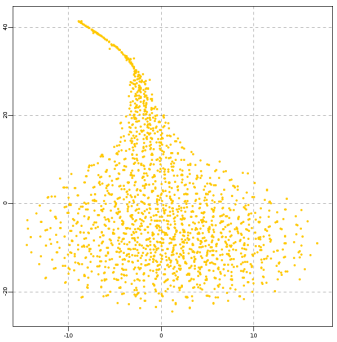

Abbildung 3.4.: Stemmer

Abbildung 3.5 zeigt das Ergebnis nach einem Clustering anhand einer Distanzmatrix basierend auf den Cosinus-Distanzen der normalisierten Vektoren. Es ist erkennbar, dass die Cluster bereits einigermaßen abgegrenzt sind und sich mit der Visualisierung decken.

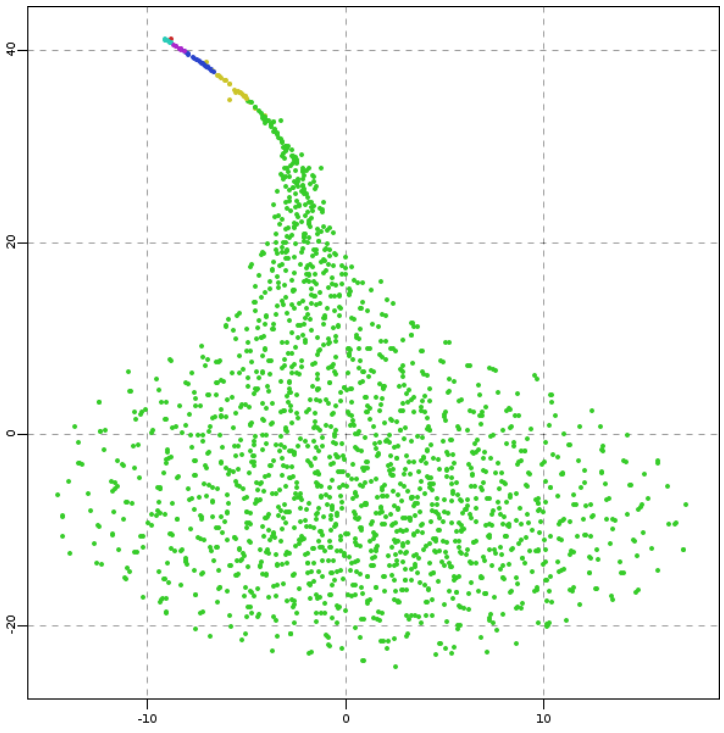

Abbildung 3.5.: Wortvektoren nach Clustering

3.2. Anwendung der Wortvektoren

Die Transformation der Rohdaten in eine Repräsentationsform besteht aus der Anwendung der Wortvektoren auf die Tasting Notes aus dem Testdatenset. Der Median der Vektoren der Wörter aus einer Tasting Note bildet nun die Repräsentation des zugehörigen Whiskys. Auf diese Weise können die Whiskys in Folge analog zu den Wortvektoren in Kapitel 3.1 verglichen und klassifiziert werden.

Die Ergebnisse dieser Anwendung zeigen die Grafiken 3.6 und 3.7. Beide zeigen die Verteilung der Whiskys nach einer Dimensionsreduktion mit t-SNE. Abbildung 3.6 zeigt farblich abgegrenzt die resultierenden Cluster aus der Berechnung einer Distanzmatrix auf Basis der Cosinus-Distanzen. Als Kriterium für die Clusterbildung wurde hier Complete Linkage verwendet. Abbildung 3.7 zeigt die resultierenden Cluster aus der gleichen Berechnung auf Basis der euklidischen Distanzen. Auch hier lässt sich bereits der Trend erkennen, dass sich die graphische Darstellung mit der Zuordnung der Cluster deckt.

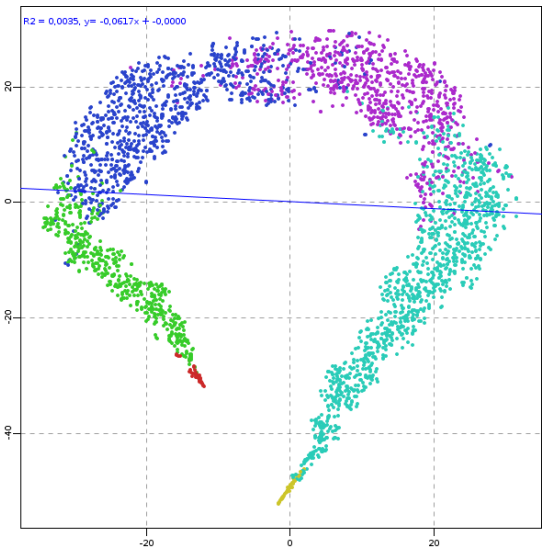

Abbildung 3.6.: Clustering anhand der Cosinus-Distanz

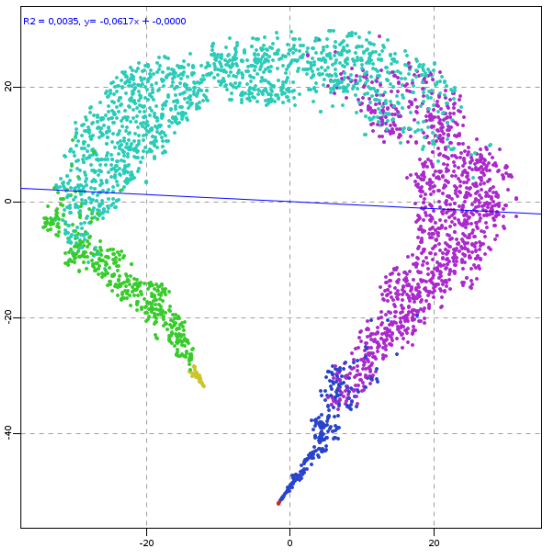

Abbildung 3.7.: Clustering anhand der euklidischen Distanz

4. Evaluierung der Versuchsergebnisse

Die erlangten Versuchsergebnisse zeigen den Einfluss der verschiedenen Vorverarbeitungsmethoden auf die Wort- und Whiskyvektoren. Besonders ist der Einfluss des Stopwordfilters zu sehen. Nicht dargestellt ist eine Unterscheidung der Ergebnisse ohne und mit Anwendung des Stemmers, da hier nur geringe Unterschiede sichtbar sind. Weiterhin ist nicht dargestellt, wie sich das Clustering ohne und mit vorheriger Normalisierung der Vektoren verhält. Die Cluster nach vorheriger Normalisierung sind abgegrenzter zueinander und optisch deutlich nachvollziehbarer.

Es zeigt sich an den Wortvektoren, dass ein größeres Trainingsset hilfreich wäre. Gut erkennbar ist jedoch bereits, dass sich Wörter, welche Geschmacksrichtungen von Whiskys beschreiben, in der oberen Linie in Abbildung 3.5 sammeln, also im Verhältnis zu anderen Wörtern nahe beieinander liegen. Eher irrelevante Wörter grenzt das Clustering deutlich ab, was am großen, grünen Cluster erkennbar ist. Diesen Eindruck bestätigt auch eine genauere Betrachtung der Cluster. Während sich vier der sechs gebildeten Cluster aus Wörtern zusammensetzen, die häufig zur Beschreibung von Whiskys verwendet werden, beinhaltet der größte Cluster vorwiegend eher irrelevante Begriffe, aber auch Geschmacksbegriffe. Dies ist ein weiteres Indiz dafür, dass das Verfahren noch verfeinert werden muss. Die Vermutung liegt nahe, dass die Geschmacksbegriffe im großen Cluster nicht häufig genug im Trainingsdatenset vorkommen. Der sechste Cluster ist zu vernachlässigen, da dieser nur einen Bindestrich als Begriff enthält. Im großen Cluster befinden sich ebenso noch Sonderzeichen und Stopwords, was zeigt, dass bei der Vorverarbeitung der Texte ebenfalls Verbesserungspotential besteht. Die resultierenden Cluster sind in Anhang A aufgeführt.

Die Anwendung der Wortvektoren auf die Whiskys zeigt, dass sich Cluster ermitteln lassen, diese für ein zufriedenstellendes Ergebnis allerdings noch deutlich abgegrenzter zueinander sein müssen. Ebenso ist erkennbar, dass die Cosinus-Distanz geeigneter als die euklidische Distanz als Distanzmaß ist, da diese im Vergleich klarere und in der Größe ausgeglichenere Cluster bildet.

5. Ansätze zur Optimierung des KDD-Prozesses

Der gewählte KDD-Prozess bietet in jedem Schritt noch erkennbare Optimierungsmöglichkeiten. Das Trainingsset kann jederzeit durch mehr Daten vergrößert werden und damit bei der Erzeugung besserer Wortvektoren helfen. Weiterhin kann eine verfeinerte Aufbereitung der Rohdaten die Vektoren verbessern. Die einzelnen Schritte wie das Stopword-Removal können beispielsweise durch bessere Stopwordlisten optimiert werden oder durch eine Positivliste ersetzt werden. Diese Maßnahme hätte zur Folge, dass nur noch Wörter berücksichtigt werden, welche auch tatsächlich aus der Domäne stammen. Gleichzeitig birgt sie aber auch das Risiko, dass einige relevante Begriffe nicht berücksichtigt werden, da sie in der Positivliste fehlen. Ein möglicher zukünftiger Versuch wäre ein Vergleich beider Wege mit anschließender Evaluation.

Für ein besseres Training der Wortvektoren bietet sich die außerdem Möglichkeit, die Rohtexte ontologiebasiert anzureichern. Das würde konkret bedeuten, die Sätze mit Oberbegriffen vorkommender Wörter anzureichern. Somit würden synonyme Unterbegriffe häufiger mit den gleichen Oberbegriffen in einem Kontext auftauchen und dadurch eher als synonym erkannt werden. Ebenso bietet sich die Möglichkeit, die Parametrisierung der einzelnen Algorithmen weiter anzupassen und den Einfluss auf die Ergebnisse zu prüfen. Weiterhin ist es möglich, andere Methoden des Clusterings und der Distanzberechnung auf ihre Eignung zu testen. Außerdem gilt es, Methoden der automatischen Cluster-Evaluierung in den Prozess einzubinden.

6. Zusammenfassung und Ausblick

Diese Arbeit beschreibt die erste Durchführung eines zuvor geplanten KDD-Prozesses mit Hilfe der Software KNIME und zugehörigen Plugins. Die Arbeit erläutert die einzelnen Schritte des KDD-Prozesses mit konkretem Bezug zum Projekt. Sämtliche Schritte des Prozesses sind einmal durchgeführt und anschließend bewertet worden. Die Ergebnisse zeigen erste Erfolge, aber auch Verbesserungsmöglichkeiten in allen Teilschritten. Die entstehenden Cluster zeigen Tendenzen einer nachvollziehbaren Aufteilung. Speziell irrelevante Begriffe werden bereits in einem Cluster zusammengefasst. So stellt sich vor allem heraus, dass eine Vergrößerung des Trainigsdatensets und eine ontologiebasierte Anreicherung der Texte erfolgsversprechende Optimierungsmöglichkeiten sind. Diese Schritte bilden einen Teil der Hauptziele der anschließenden Masterthesis beziehungsweise der Vorbereitung dieser.

A. Ergebnis des Clusterings der Wortvektoren

Cluster 0 -

Cluster 1 amber, award, banana, berri, bitter, black, cherri, cinnamon, classic, cream, dri, follow, fresh, gentl, ginger, glass, gold, herbal, initi, mix, nice, nut, polish, spice, strong, sugar, time, toffe, yellow, zest

Cluster 2 -year-old, –, …, 'green', ', "the, ", £, a, aberlour, absolut, abund, abv, accent, accentu, accompani, aceton, acid, acrid, act, activ, actual, ad, add, addit, adelphi, afford, afor, aftertast, ag, agav, aggress, ago, ah, ahh, air, airi, akin, al, alarm, albeit, alberta, alcohol, aliv, allow, allspic, almond, alongsid, alpin, altern, altogeth, am, amaretto, amaz, amazingli, america, american, amidst, amongst, amontillado, amount, amplifi, amrut, ancient, anim, anis, anise, anniversari, anonym, anticip, antiqu, antisept, anywai, apart, aperitif, appar, appeal, appear, approach, apricot, ardbeg, ardmor, armagnac, aroma, aromat, arran, arriv, artifici, ash, ashi, asid, aspect, assert, associ, astring, atlant, attack, attent, attract, attribut, auchentoshan, auster, autumn, avail, awesom, b', babi, backdrop, background, bacon, bad, bag, bai, bake, balanc, ball, balsam, balveni, bandag, bang, bar, barbecu, bare, bark, barlei, barn, barrel, base, basil, basket, batch, battl, bbq, beach, bean, bear, beast, beauti, beautifulli, becom, bee, beef, beer, beeswax, begin, believ, below, ben, beneath, benefit, benriach, benromach, berryish, beyond, bigger, bilg, bind, biscuit, biscuiti, bite, blackberri, blackcurr, bland, blast, blend, blender, blind, blood, bloom, blossom, blue, blueberri, blush, board, bodi, bog, boil, bold, bomb, bone, bonfir, bonu, book, boot, bore, bottl, bottler, bouquet, bourbon, bourboni, bovril, bowl, bowmor, box, brace, bracken, brambl, brand, brandi, brass, brasso, bread, breadi, break, breakfast, breath, breez, brief, briefli, bright, brilliant, brine, bring, brini, brioch, brisk, british, brittl, broad, bronz, brora, brought, brown, bruichladdich, brule, brulé, brulè, brûlée, brutal, bt, bugatti, bui, build, bunnahabhain, burgundi, burn, burnt, burst, busi, butt, butter, butteri, butterscotch, c, cabinet, cacao, cadenhead, cake, calgari, call, calm, campfir, camphor, can, can't, canada, canadian, candi, candl, cane, cant, cantaloup, caol, car, caramelis, carat, card, cardamom,

cardboard, cardemom, care, carri, cashew, cask, cask-driven, cassi, catch, cedar, celebr, celeri, cellar, center, centr, cereal, chalk, chalki, challeng, chamomil, chanc, chang, char, charact, characterist, charcoal, charg, charm, check, chees, chemic, chestnut, chew, chewi, chicori, chieftain', childhood, chili, chill, chines, chip, chocolate-cov, chocolati, christma, cider, cigar, cigarett, citric, citrusi, civilis, cl, clan, class, classi, clean, clear, cling, clinic, cloi, close, cloth, clove, clover, clynelish, coal, coast, coastal, coat, cocktail, cocoa, coconut, coffe, cognac, coher, cold, collect, color, combin, come, comment, compact, compani, compar, compass, competit, complet, complex, complic, compliment, compon, compos, compot, compris, concentr, conclud, confus, consid, consider, consist, contain, continu, contrast, control, cook, cooki, cool, copper, copperi, core, coriand, cork, corn, correct, cotton, cough, countri, countrysid, coupl, cours, cover, cow, crack, cracker, craft, cranberri, crash, crazi, creamier, creat, creep, creme, crème, creosot, crisp, crispi, critic, cross, crumbl, crunchi, crust, crystallis, ct, cucumb, cupboard, cure, curiou, currant, current, curri, custard, custom, cut, d, d', dai, dalmor, damp, damson, danc, danish, darker, darkli, dash, date, david, de, deal, dear, decad, decent, decid, decidedli, deep, deepen, deeper, defin, definit, delic, delici, delight, delightfulli, deliv, deliveri, demand, demerara, demonstr, dens, departur, depth, describ, descriptor, deserv, desper, despit, dessert, detail, detect, deterg, detract, develop, dewar, di, diesel, differ, difficult, diffus, dilut, dimens, diminish, dinghi, dinner, direct, dirt, dirti, disappear, disappoint, discreet, dissip, distanc, distant, distil, distilleri, distillery', distinct, distinctli, disturb, dollop, domin, don't, dont, dose, doubl, dough, dougla, dour, drag, dram, dramat, drier, drift, drink, drinkabl, drizzl, drop, dryish, dryness, due, dull, duncan, dunnag, duski, dust, dusti, duthi, earlier, earth, earthi, earthier, eas, easi, easili, easy-go, edg, edgi, edinburgh, edit, edradour, effect, egg, eight, elder, eleg, element, ellen, els, ember, emerg, emphasi, empti, encount, energet, energi, english, enhanc, enjoi, enjoy, enorm, enthusiast, entir, entri, equal, especi, espresso, esteri, etc, etcetera, eucalypti, eucalyptu, european, even, eventu, everlast, everydai, evid, evolv, ex, ex-bourbon, exactli, examin, exampl, exceedingli, excel, except, exception, exchang, excit, exclus, exhibit, exist, exot, expand, expect, expens, experi, experienc, explod, explor, explos, express, extra, extraordinari, extrem, ey, facet, fade, faint, faintest, faintli, fair, fairli, fall, fame, famili, familiar, famou, fan, fantast, farm, farmi, farmyard, fascin, fashion, fat, favourit, featur, feel, fennel, festiv, field, fierc, fieri, fig, fight, fill, filter, filtrat, final, fine, finest, fino, fire, fireplac, firm, first-fil, fish, fishi, fit, five, fizzi, flake, flash, flat, flavor, flavour, flawless, floor, flora, floss, flow, flower, floweri, focus, food, fool, forc, fore, forest, forev, forget, form, forth, fortun, forward, found, fragil, fragranc, fragrant, frankli, free, french, fresher, freshli, friend, front, fruitcak, fruitier,

fruity, fry', fudg, full-bodi, fuller, fun, funki, furnitur, fyne, gain, galor, garden, gauz, gener, gentler, gentli, german, get, giant, gin, gingerbread, give, glaze, glen, glencairn, glenfarcla, glenfiddich, glenlivet, glenmorangi, glenroth, gloriou, glove, glow, glue, gm, goe, golden, gone, good, gooseberri, gordon, gradual, graham, grain, graini, grand, grant, grant', grape, grapefruit, grappa, grass, grassi, great, greatli, green, greet, grei, grill, grip, ground, grow, guava, guess, guis, gum, gun, gunpowd, hai, hair, half, ham, hand, handl, hang, happen, happi, happili, harbour, hard, hardbread, hardli, harmoni, harsh, hazelburn, hazelnut, head, headi, heart, heat, heather, heatheri, heavi, heavier, heavili, hefti, held, help, herb, herbac, heritag, hessian, hidden, hide, highland, highli, highlight, hike, hill, hit, hm, hogshead, hold, home, honest, honeycomb, honeysuckl, hope, hors, hospit, hot, hour, hous, hover, hp, httpwwwlfwcouk, hue, huge, hurt, i'd, i'm, i'v, ic, idea, identifi, ila, imag, imagin, immatur, immedi, immens, impact, import, improv, incens, includ, increas, increasingli, incredibli, inde, independ, indian, indic, individu, industri, influenc, inform, infus, insignific, instead, instrument, integr, intellig, intens, intensifi, intermingl, interplai, intertwin, intrigu, introduc, intrus, invit, involv, iodin, irish, isl, islai, island, it', italian, jam, jammi, japanes, jasmin, jelli, jetti, jim, join, jolli, journei, juic, juici, jump, junip, jura, kei, kelch, kensington, kept, kick, kilchoman, kiln, kindli, king, kipper, kiwi, la, label, lace, lack, laddi, lagavulin, laid-back, laphroaig, lash, last, late, latter, lavend, lavendar, layer, lead, leaf, leafi, lean, leather, leatheri, leav, ledaig, left, leg, legendari, lemongrass, lemoni, length, level, licoric, life, lift, lighten, lighter, lightest, lightli, like, likewis, lime, limit, line, linen, linger, linkwood, linse, liqueur, liquid, liquoric, lit, live, liveli, ll, load, loaf, locat, loch, log, lolli, lomond, longmorn, look, lose, lost, love, lover, low, lower, lowland, lumber, lurk, lusciou, luxuri, lyche, m, macallan, mace, macphail, madeira, magic, magnific, mahogani, main, mainli, make, malti, malty, manag, mandarin, mango, manner, mapl, marin, maritim, mark, market, marmalad, marmit, marri, marshmallow, marvel, marzipan, mash, mask, mass, massiv, master, match, matter, matur, mayb, meadow, mean, measur, meat, meati, medal, medicin, meet, mellow, melon, melt, memori, menthol, mention, mere, metal, mid, mid-pal, middl, mild, mildli, milk, milki, mill, mind, miner, mingl, minim, mint, minti, minut, mirror, miss, mixtur, mma, mocha, moder, molass, moment, monei, monster, month, moor, morai, moreish, morn, mortlach, moss, mouth, mouth-coat, mouth-feel, mouthfeel, move, mr, muesli, mull, multi, munich, murrai, mushroom, music, mustard, musti, mute, n't, na, nai, name, natur, near, nearli, neat, neither, never-end, nevi, newli, nicer, night, no, non-descript, none, nor, normal, north, notic, nougat, nt, nuanc, nutti, oaki, oat, oatmeal, ob, obscur, obviou, obvious, ocean, odd, oddish, oddli, offer, offici, offnot, oh, oil, oili, ok, old, old-fashion,

oliv, oloroso, on, one-dimension, onto, ooz, opinion, opportun, opposit, orangei, orchard, organ, orient, origin, otherwis, ought, outdoor, outstand, overli, overpow, overrip, overt, overton, overwhelm, own, owner, oxid, oxidis, oyster, pack, packag, paint, pale, papaya, paper, park, parma, parti, particular, particularli, pass, passion, past, pastil, pastri, patienc, pea, peach, peachi, peak, peanut, pear, peati, pecan, pedro, peel, penetr, peopl, peppercorn, pepperi, peppermint, peppery, perfect, perfectli, perfum, period, persist, person, petal, phenol, pick, pickl, pictur, pie, piec, pier, pine, pineappl, pink, pipe, pizza, plai, plain, plasticin, pleas, pleasant, pleasantli, pleasur, plenti, plu, plum, plump, poach, polit, pollen, pool, pop, popcorn, pork, porridg, port, posit, possibli, pot, pot-pourri, potato, potent, pour, powder, powderi, power, ppm, precis, predomin, prefer, premium, prepar, presenc, preserv, press, pretti, previou, previous, price, prickl, prickli, primarili, prior, pristin, privat, probabl, process, produc, product, profil, progress, promin, promis, pronounc, proof, prove, provid, prune, pud, puff, pull, pultenei, pump, punch, punchi, pungent, purchas, pure, push, put-togeth, px, qualiti, quarter, quick, quickli, quiet, quinc, quit, rais, rancio, rang, rapidli, rare, raspberri, raspi, rate, rattrai, raw, re, reach, read, real, reappear, reason, reced, receiv, recent, recommend, red, redcurr, reddish, reduc, reduct, reek, refil, refin, reflect, refresh, regard, region, rel, relax, releas, remain, remark, rememb, remind, reminisc, remov, repeat, replac, repres, requir, resembl, reserv, resin, rest, restless, restrain, result, retail, retain, retast, retronas, return, reveal, review, revisit, reward, rhubarb, rice, richer, richli, ride, rind, ripe, rise, road, roast, robust, rock, roll, roof, root, rope, rose, rosebank, rosemari, rough, round, rounder, rowan, rubber, rubberi, rum, run, rush, rye, sadli, sage, sai, salad, sale, salin, salmiac, salt, salti, sampl, sand, sandalwood, sandi, sangria, sap, sappi, satisfi, sauc, sauna, sautern, save, savori, savouri, sawdust, sawn, scent, scorch, score, scotch, scotland, scotland', scottish, sea, sea-air, seal, search, seashor, season, seat, seawe, seduct, seed, seek, seen, select, sell, sens, sensat, seri, seriou, serious, serv, set, settl, sevil, shade, shake, shame, shape, share, sharp, sharper, shave, shed, shell, sherbet, shift, shine, shini, shock, shoe, shop, short, shy, sight, sign, signatur, signific, significantli, silki, silver, similar, simpl, simpler, simpli, sing, singl, sip, sippabl, sit, six, skin, slice, slighli, slight, slightest, slighti, sligthli, slow, slowli, smack, smell, smokei, smoker, smokier, smolder, smoother, smoulder, smw, snap, soak, soap, soapi, societi, soda, soft, soften, softer, softli, soi, sold, solid, solvent, somehow, sometim, somewhat, soon, soot, sooth, sooti, sophist, sorri, sort, sound, soup, sour, sourish, sourli, speak, spearmint, special, specif, spend, spent, speysid, spicier, spicy, spiegelau, spirit, spiriti, splash, split, spread, spring, spring-tim, springbank, sprinkl, spritzi, st, stabl, stage, stai, stale, stand, standard, star, start, statement, steadili, steal, steam, steep, stem, step, stew,

stick, sticki, sting, stingi, stock, stone, stop, store, straight, straightforward, strang, straw, strawberri, strength, strike, stroke, stronger, strongli, struck, structur, struggl, stuff, stun, style, subdu, substanc, subtl, subtli, success, succul, sudden, suddenli, suffici, sugari, suggest, suit, sulfur, sulphur, sulphuri, sultana, summari, summer, sun, sunni, super, superb, suppl, sure, surfac, surpris, surprisingli, surupi, suryp, suspect, swallow, sweeten, sweeter, sweetish, sweetli, swim, syrup, syrupi, syruppi, tabl, tad, tail, taiwan, take, talisk, talk, tame, tang, tangerin, tangi, tannic, tannin, tar, tarri, tart, tast, tastewis, tasti, taylor, tea, teak, tell, temperatur, tend, tendenc, term, textur, th, thank, thankfulli, theme, there', thick, thicker, thin, think, thinner, third, thoroughli, throughout, throw, thrown, tide, tight, tin, ting, tingl, tingli, tini, tip, tire, toast, tobacco, tobermori, tomato, ton, tone, tongu, top, toss, total, touch, tough, toward, trace, trade, tradit, trait, transform, translat, transport, travel, treacl, treat, tree, tri, tropic, true, truli, try, tube, tullibardin, turkish, tussl, twig, twist, type, typic, uk, ultim, unbalanc, uncompl, unctuou, undercurr, underli, underneath, underst, underton, undertow, undilut, unexpect, unfortun, unhappi, uniqu, unknown, unlik, unlit, unpeat, unpleas, unsweeten, unusu, urquhart, us, usa, usual, vagu, valu, vanillaish, vanish, variant, varieti, variou, varnish, vat, ve, veget, vein, velveti, verit, version, vetiv, vibranc, vibrant, vine, vinegar, vintag, violet, virgin, visit, vomit, waft, wait, walk, walker, walnut, warehous, warm, warmer, warmth, wash, wateri, wave, wax, waxi, we'r, weak, weaken, weaker, weakish, weather, websit, wee, week, weetabix, weight, weird, weirdli, welcom, well-balanc, well-integr, welli, wet, what', whatev, wheat, whelm, wherea, whiff, whilst, whip, whiskei, white, whoa, wi, wild, william, win, wine, winei, winter, wise, wish, wonder, wonderfulli, wont, wooden, woodi, woodsmok, woody, word, world, worth, worthi, wow, wrap, wrong, ximenez, y, yard, ye, yeasti, yo, you'd, you'll, you'r, youngish, youth, yummi, zesti

Cluster 3 colour, finish, fruit, fruiti, hint, malt, nose, overal, peat, sweet

Cluster 4 appl, bit, caramel, chocol, citru, creami, dark, floral, impress, lemon, littl, lot, medium, note, nutmeg, oak, pepper, raisin, sherri, smoki, smooth, vanilla, water, whiski, wood

Cluster 5 dry, honei, light, orang, palat, rich, slightli, smoke, spici

Literaturverzeichnis

[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

[Deeplearning4j Development Team 2017] Deeplearning4j Development Team: *Deeplearning4j: Open-source distributed deep learning for the JVM*. http://deeplearning4j.org. 2017. – Letzter Zugriff am 04.02.2018

[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

[van der Maaten 2017] Maaten, Laurens van der: *BHTSNE*. https://github.com/lvdmaaten/bhtsne. 2017. – Letzter Zugriff am 04.02.2018

[Mikolov u. a. 2013] 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

[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

[Scotchwhisky.com 2016] Scotchwhisky.com: *Scotchwhisky.com*. https://scotchwhisky.com. 2016. – Letzter Zugriff am 04.02.2018

[Ter Heide 2017] Ter Heide, Dominiek: *Python BHTSNE*. https://github.com/dominiek/python-bhtsne. 2017. – Letzter Zugriff am 04.02.2018

[WhiskyIntelligence 2017] WhiskyIntelligence: *Whiskyintelligence.com*. http://whiskyintelligence.com/. 2017. – Letzter Zugriff am 04.02.2018

[WhiskyMagazine 2016] WhiskyMagazine: *Whisky Magazine*. https://www.whiskymag.com. 2016. – Letzter Zugriff am 04.02.2018

[WhiskyMonitor 2017] WhiskyMonitor: *Whisky Monitor*. https://www.whisky-monitor.com. 2017. – Letzter Zugriff am 04.02.2018