UNIVERSITÄT LEIPZIG

Information Retrieval - Praktikum

Wine Search

Timo Lehmann, Maik Bachmann, Martin Lorenz

beaufsichtigt von:

Junior-Prof. Dr. Martin Potthast

11-08-2019

Inhalt

1. Einleitung

Der folgende Bericht soll unser Vorgehen und unsere Entscheidungen für das Projekt Wine-Search erläutern. Wine-Search ist eine Suchmaschine, welche im Rahmen des Moduls *Information Retrieval* entwickelt wurde. Das Ziel des Projekts war es, an einer selbstgewählten Domäne (Weinrezensionen) die aus der Vorlesung vermittelten Inhalte selbst praktisch auszuprobieren und zu vertiefen. Die Schwerpunktsetzung war uns relativ freigestellt, aber die Anwendung sollte nach Aufgabenstellung auf einem Server laufen, mittels eines für Suchmaschinen üblichen Web-Interfaces bedienbar sein, einen Evaluierungsmodus besitzen und die standard Funktionsweise einer Suchmaschine haben. Letzteres bedeutet konkret, die Daten zu Indexieren, in diesen Daten zu suchen und alle Anfragen mit den dazugehörenden relevanten Informationen zu loggen.

2. Motivation

Das Thema Wein und insbesondere auch das Trinken von Wein beschäftigt die Menschen schon seit mehreren Tausenden von Jahren[1]. Im Laufe der Zeit wurden viele neue verschieden Rebsorten gezüchtet, welche heute abhängig von ihren ganz individuellen Anforderungen weltweit angebaut werden. Diese Vielfalt macht es einem schwer, den Überblick zu behalten bzw. einen neuen Wein für sich zu entdecken, der einem schmeckt. Allerdings gibt es Portale im Internet, auf denen anerkannte Wein-Rezensenten ihre Beurteilungen für die verschiedensten Rot- und Weissweine aus aller Welt veröffentlichen. Eine solche Seite ist beispielsweise WineEnthusiast, von welcher unser Datensatz stammt. Mittels Wine-Search kann dieser nun vom Nutzer durchsucht werden, wobei nicht nur Eigenschaften wie Farbe, Aroma, Rebsorte und Herkunftsland, sondern auch Eigenschaften wie der Preis berücksichtigt werden.

3. Der Datensatz

Der Datensatz mit den Wein-Rezensionen stammt von der Plattform Kaggle, wobei die Reviews von der Seite winemag.com gecrawlt wurden. Es liegen fast 130.000 Rezensionen im CSV-Format vor, wobei die neben der Rezension an sich noch folgende Felder für fast alle Rezensionen gegeben sind: Ursprungsland des Weins, die Kellerei, der Titel, welcher oft auch den Jahrgang enthält, der Preis und die Provinz bzw. Region aus der der Wein im Ursprungsland stammt. Es gibt auch Attribute, die weniger relevant sind. Dies wären der Tester bzw. der Rezensent und die von ihm subjektiv vergebene Anzahl an Punkten für einen Wein. Im folgenden wird darauf eingegangen, wie wir zu dieser Schlussfolgerung kamen.

3.1. Analyse des Datensatzes

Das Ziel der Analyse bestand darin, sich einen Überblick über den Datensatz zu verschaffen, um im Nachhinein Queries zu formulieren, die den Datensatz größtenteils abdecken. Es sollten also auch Themen gefunden werden, die nicht auf den ersten Blick offensichtlich waren.
Dazu wurde im ersten Schritt der Datensatz „per Hand“ analysiert, d.h. es wurden Auswertungsmöglichkeiten genutzt, die einem ein Standard Tabellenprogramm bietet. Ein Beispiel dafür ist die Anzahl unterschiedlicher Werte. So wurde mittels dieser Vorgehensweise ermittelt, dass die 130.000 Rezensionen von nur 19 verschiedene Rezensenten erstellt wurden und dass fast 20% aller Rezensionen auf eine Person zurückgehen. Da somit fast 26.000 Rezensionen von einer Person stammen, ist das Suchen nach Rezensionen von bestimmten Personen eher irrelevant.
Des Weiteren viel auf, dass es zwar eine subjektive Anzahl vergebener Punkte auf einer Skala von 0 bis 100 gab, dass nur Punkte im Bereich von 80 bis 100 vergeben wurden. Wobei sich hier 90% der Rezensionen im Bereich von 84 bis 94 Punkten bewegen.
Ein weiterer Punkt, der uns zu einer tiefgründigeren Analyse veranlasste war, dass die Rezensionen alle recht ähnlich waren. Der Großteil beschreibt Geschmacksnuancen mit sehr ähnlichen Wörtern, was natürlich ein konkretes Suchen erschwert. Ein weiterer auffallender Punkt

war, dass sehr viele Adjektive verwendet wurden. Deshalb soll der nächste Teil dieses Kapitels beschreiben, wie mittels Methoden des Information Retrieval der Datensatz analysiert wurde.

3.2. Analyse der Rezensionen

Da hauptsächlich die Rezensionen durchsucht werden sollen, werden diese hier genauer betrachtet. Zu allererst wurde die durchschnittliche Rezensionslänge bestimmt. Dies wurde während des Indexierungsprozesses erledigt. Das Ergebnis war eine durchschnittliche Länge von 40 Wort, was für eine Volltextsuche recht kurz ist, aber noch toleriert werden kann.

3.2.1. Verwendete Adjektive

Da uns die Analyse bei der Formulierung der Queries helfen soll, wurde im nächsten Schritt die Wortwahl genauer betrachtet, konkret, die Häufigkeit der verwendeten Adjektive. Dies wurde mittels Apache OpenNLP und Apache Lucene realisiert. Dafür wurde ein Modul erstellt, welches mittels Part-Of-Speech-Tagging alle verschiedenen Adjektive zählt. Das Ergebnis dabei war, dass es ca. 11.000 verschiedene Adjektive gibt. Wobei ungefähr die Hälfte nur ein einziges mal vorkommt und ca. 1% mehr als 1000 mal vorkommen.

Wort	Häufigkeit
ripe	24975
black	24325
red	17177
fresh	16909
rich	16399
dry	13999
soft	12196
white	11993
bright	10870
dark	10713

Wort	Häufigkeit
good	9404
green	8585
crisp	8259
smooth	6523
fine	5843
creamy	5605
tannic	5522
clean	5486
juicy	5435
full-bodied	5222

Wort	Häufigkeit
dense	5002
ready	4984
great	4934
elegant	4912
earthy	4653
savory	4652
tight	4512
delicious	4507
complex	4434
concentrated	4243

Abb 1: Die 30 häufigsten Adjektive

Für die Query-Formulierung sind nicht alle Adjektive relevant. Man wird wohl kaum eine Anfrage formulieren, die explizit Wörter wie *ripe, good, fine* oder *delicious* enthält. Dies sind eher implizierte Vorraussetzungen des Suchenden. Für Anfragen sind eher Worte wie *dry, red* oder *white* interessant, allerdings stellt dies keinen neuen Erkenntnis gewinn dar. Worte wie soft, *smooth, creamy, full-bodied* oder *concentrated*, sind wohl eher der begeisterten Ausdrucksweise des Rezensenten zuzuschreiben, als das es wirklich relevante Suchbegriffe wären.
Dennoch wurde mit dieser Auswertung und etwas Suchaufwand das Themengebiet „Bio“ herausgearbeitet und enthält Wörter wie *organic, biodynamic, regional, demeter-certified* und *certified-organic*. Dies ist jetzt eins von potentiell vielen Themen.

3.2.2. Clustering mittels LSI

Um einen besseren Überblick über die verschiedenen Themen der Rezensionen zu bekommen, wurde mittels des LSI-Verfahrens eine Übersicht der Themen anhand der verwendeten Wörter erstellt. Bei einem ersten Durchgang erhielten wir das Ergebnis: 4 verschiedene Themen bei einer maximalen Kohärenz von 0,468. Man hat also selbst bei vier großen Themenkomplexen noch eine sehr hohe Überschneidung, was auch durch *Abb. 1* bestätigt wird. Also wurden im zweiten Schritt die häufigsten Wörter wie *wine, palate* und auch die 120 häufigsten Adjektive entfernt.

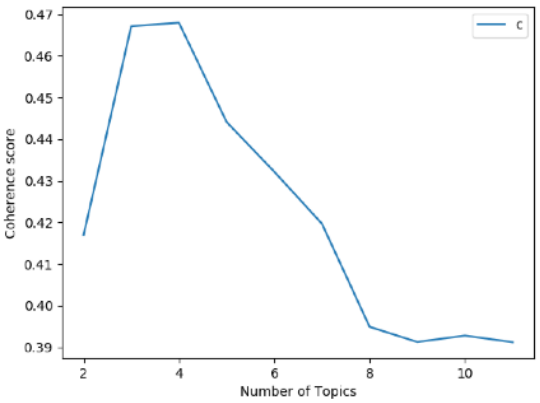

Abb. 2: *Kohärenz Wert in Abhängigkeit der Themenanzahl bei Durchgang 1*

1: wine, flavor, fruit, aroma, acid, finish, palate, drink, tannin, cherry, black, ripe, dry, spice, note, rich, red, fresh, berri, show

2: wine, aroma, palat, finish, flavor, cherri, note, fruit, age, nose, black, drink, rich, berri, plum, spice, offer, acid, fruiti, ripe

3: black, tannin, flavor, cherri, appl, acid, finish, fresh, citru, lemon, crisp, peach, pear, white, dark, red, spice, cabernet, blackberri, green

4: fruit, flavor, palat, wine, cherri, fresh, blackberri, white, note, acid, nose, app, offer, oak, citru, cabernet, alongsid, sweet, chocol, hint

Abb. 3: *Häufigste Wortstämme in den Themen und ihre Titel, 1: fruchtige trockene Rotweine, 2: würzige Rotweine mit pikanter Säure, 3: Rezensionen mit vielen Geschmacksnuancen, 4: fruchtige Weißweine*

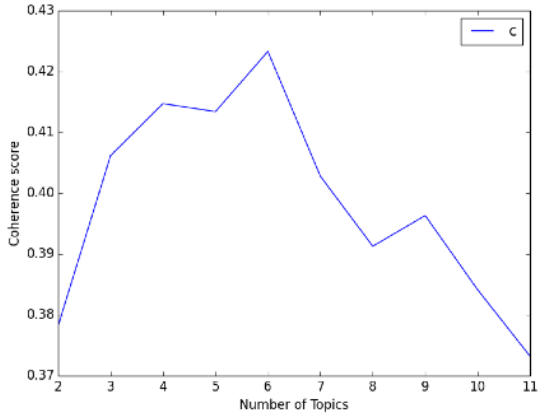

Abb. 4: *Kohärenz Wert in Abhängigkeit der Themenanzahl bei Durchgang 2*

1: fruit, flavor, aroma, finish, acid, drink, tannin, berri, oak, note, plum, show, full, nose, blackberri, offer, blend, sweet, age, appl

2: fruit, flavor, finish, aroma, drink, age, acid, tannin, plum, berri, give, nose, feel, note, wood, appl, well, charact, structur, year

3: tannin, flavor, aroma, acid, berri, appl, cabernet, plum, blackberri, fruit, lemon, offer, peach, pepper, alongsid, blend, drink, licoric, lime, sweet

4: fruit, drink, acid, finish, aroma, tannin, flavor, fruiti, age, give, year, textur, note, well, blackberri, nose, charact, feel, full, structur

5: flavor, acid, aroma, appl, finish, offer, cabernet, lemon, blackberri, nose, note, fruit, peach, hint, alongsid, miner, oak, blend, flower, lime

6: aroma, finish, cabernet, note, blend, flavor, berri, sauvignon, oak, nose, merlot, show, fruit, blackberri, franc, lemon, petit, sweet, appl, acid

Abb. 5: *Häufigste Wortstämme in den Themen bei Durchgang 2 und ihre Titel,*
1: beerige Weine, 2: trinke jetzt / ab Jahr 3: tanninreiche Weine mit leichter Säure aus Rebsorte Cabernet 4: Textur hat noch Potenzial, deshalb noch bis zu bestimmtem Jahr liegen lassen 5: Wein aus dem Eichenfass 6: Cabernet Sauvignon und C. Franc

Im zweiten Durchgang sah das Ergebnis etwas anders aus. Jetzt sind sechs Themenbereiche erkennbar, welche sich allerdings noch stärker als in Durchlauf 1 überschneiden, was auch an einem Kohärenzwert von 0,422 erkennbar ist. Des Weiteren wurde das LSI-Verfahren auf die Wortstämme angewandt, so dass wirklich Themen erkennbar sind. Allerdings erschweren diese durch den Porter-Stemmer erstellten Stämme etwas die Auswertung. Dennoch wurde versucht, sinnvolle Themenkategorien (*Abb. 5*) zu finden, tiefere Einblicke in den Datensatz bzw. das Erkennen von ganz neuen Themen wie dem *Bio*-Thema sind allerdings ausgeblieben.
Trotzdem stellen beide Auswertungen einen guten Ausgangspunkt für die Query-Formulierung dar, da Themenkomplexe erkennbar sind, auf welche mit gezielter Query-Formulierung zugegriffen werden kann. Gut geeignet erscheinen *fruchtige Weißweine, fruchtige trockene Rotweine, würzige Rotweine* oder auch *Weine zum jetzt / später trinken*.

4. Quellen

1 Patrick McGovern et al.: *„Early Neolithic wine of Georgia in the South Caucasus“*