Projektarbeit

Big Data

Matthias Körschens, Kevin Reinke

30. Januar 2017

Inhaltsverzeichnis

0.1 Einleitung

Ziel der Projektarbeit war es, die im Rahmen der Vorlesung „Big Data“erworbenen Fähigkeiten praktisch umzusetzten. Hierfür konnten wir einen Amazonreview-Datensatz der Stanford University nutzen. Ein Dank geht an dieser Stelle an Julian McAuley, der so freundlich war uns diesen Datensatz auf Nachfrage zu verfügung zu stellen. Zur bearbeitung der Daten haben wir Apache Spark, sowie den Clusterrechner der FSU-Jena benutzt. Fragestellung unserer Bearbeitung war das erkennen von Muster in den Daten. Im speziellen haben wir uns auf die Verteilung der Reviews und das Bewertungsverhalten der Nutzer konzentriert.

0.2 Datensatz

Der in dieser Arbeit verwendete Datensatz besteht aus Reviews des Onlinehändlers Amazon aus der Zeitspanne von Mai 1996 bis Juni 2014. Es wird hier ein spezieller Teildatensatz verwendet, welcher lediglich Reviews aus der Kategorie Elektronik enthält. Dieser hat mit 7,8 Millionen Reviews von etwa 4,2 Millionen Nutzern eine Größe von 4,7 Gigabyte. Er ist vollständig im JSON-Format gehalten. Ein Beispielreview ist nachfolgend dargestellt:


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
     piano. He is having a wonderful time playing these old
     hymns. The music is at times hard to read because we think
      the book was published for singing from more than playing
      from. Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
```

```
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Die Datensätze bestehen aus der ID des Reviewers (reviewerID), der Produkt-ID (asin), dem Namen des Reviewers (reviewerName), einer Liste mit zwei numerischen Werten, die die hilfreich-Bewertungen repräsentieren (helpful), dem Inhalt der Review (reviewText), der Produktbewertung (overall), der Überschrift bzw. Zusammenfassung des Reviews (summary), dem Datum der Review (reviewTime) und dem Datum im Unix-Format (unixReviewTime). Aufgrund der gegebenen Reviewer-ID lassen sich so auch neben den Reviews Rückschlüsse über die einzelnen Reviewer ziehen.
Der Datensatz ist verfügbar unter `http://jmcauley.ucsd.edu/data/amazon/`.

0.3 Häufigkeitsverteilung der Reviews

Eine Fragestellung ist die Verteilung der Reviews auf die Nutzer. Hierfür haben wir die Reviews mit gleicher ReviewerID aufaggregiert und die Anzahl der Reviews, die ein Nutzer abgegeben hat, dem Nutzer zugeordnet. Dabei ist aufgefallen, dass die erfassten Nutzer eher wenig bewerten. Beispielsweisen gibt es etwa 2,88 von 4,2 Millionen Nutzern die nur ein einziges Review abgaben.In Abbildung 0.1 ist die Anzahl der Nutzer, die eine spezifische Anzahl an Reviews abgegeben haben, dargestellt . Für eine bessere Visualisierung wurden beide Achsen logarithmiert.

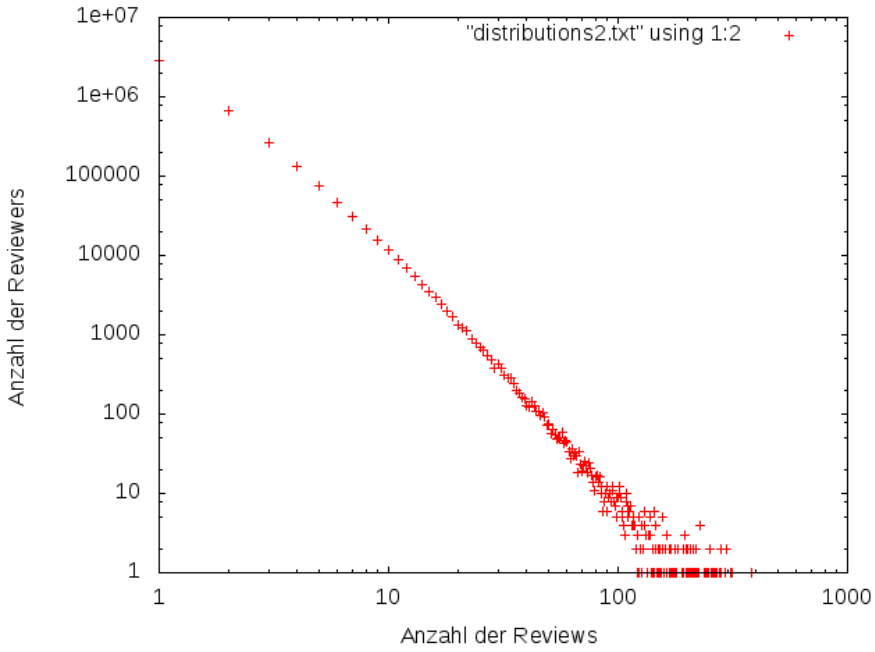

Abbildung 0.1: Reviewverteilung

0.4 Ähnlichkeitsbeziehungen

Um die Nutzer untereinander in Beziehung zu setzten, mussten wir jeden Reviewer ein Merkmalsvektor zuordnen. Diesen haben wir über arithmetische Mittelung der vom Nutzer abgegebenen Reviews bestimmt. Die Merkmale sind dabei:

- Wortanzahl des Bewertungstextes
- Zeichenanzahl des Bewertungstextes
- Zeitstempel des Reviews
- Zeichenanzahl des Reviewtitels
- wie hilfreich das Review war
- Reviewnote

Agglomeratives clustern, welches ein Distanzmaß zwischen allen Reviewern benutzt, haben wir wieder verworfen, da hierfür ein quadratischer Aufwand in Anzahl der Reviewer anfallen würde.