seit 1558

Big Data

Projekt: Amazon-Reviews

Matthias Körschens & Kevin Reinke

Friedrich-Schiller-Universität Jena

January 31, 2017

Inhalt

1 Zielsetzung

2 Datensatz

- Allgemeines
- Beispiel

3 Strukturen

- Reviewverteilung nach Reviewanzahl
- Reviewverteilung nach Textlänge

4 Clustern

Ziel

- Explorative Datenanalyse von Amazon-Reviews
- Im Speziellen:
  - Verteilung der Reviews
  - Bewertungsverhalten der Nutzer
  - Clustern von Nutzern

Allgemeines

- AmazonReviews von `http://jmcauley.ucsd.edu/data/amazon/`
- Zeitraum: Mai 1996 bis Juni 2014
- Teildatensatz: Elektronik
- 4,7 GB
- 7,8 Millionen Reviews und etwa 4,2 Millionen Nutzern
- Format: JSON

Beispiel


```
{
  "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"
}
```

Wie sind die Reviews verteilt?

- Den Nutzern wurden ihre eigenen Reviews zugeordnet
- Viele Nutzer bewerten wenig
- 2,88 von 4,2 Millionen Nutzer haben nur ein Review
- Zusammenhang exponentiell (Zipfsches Gesetz)

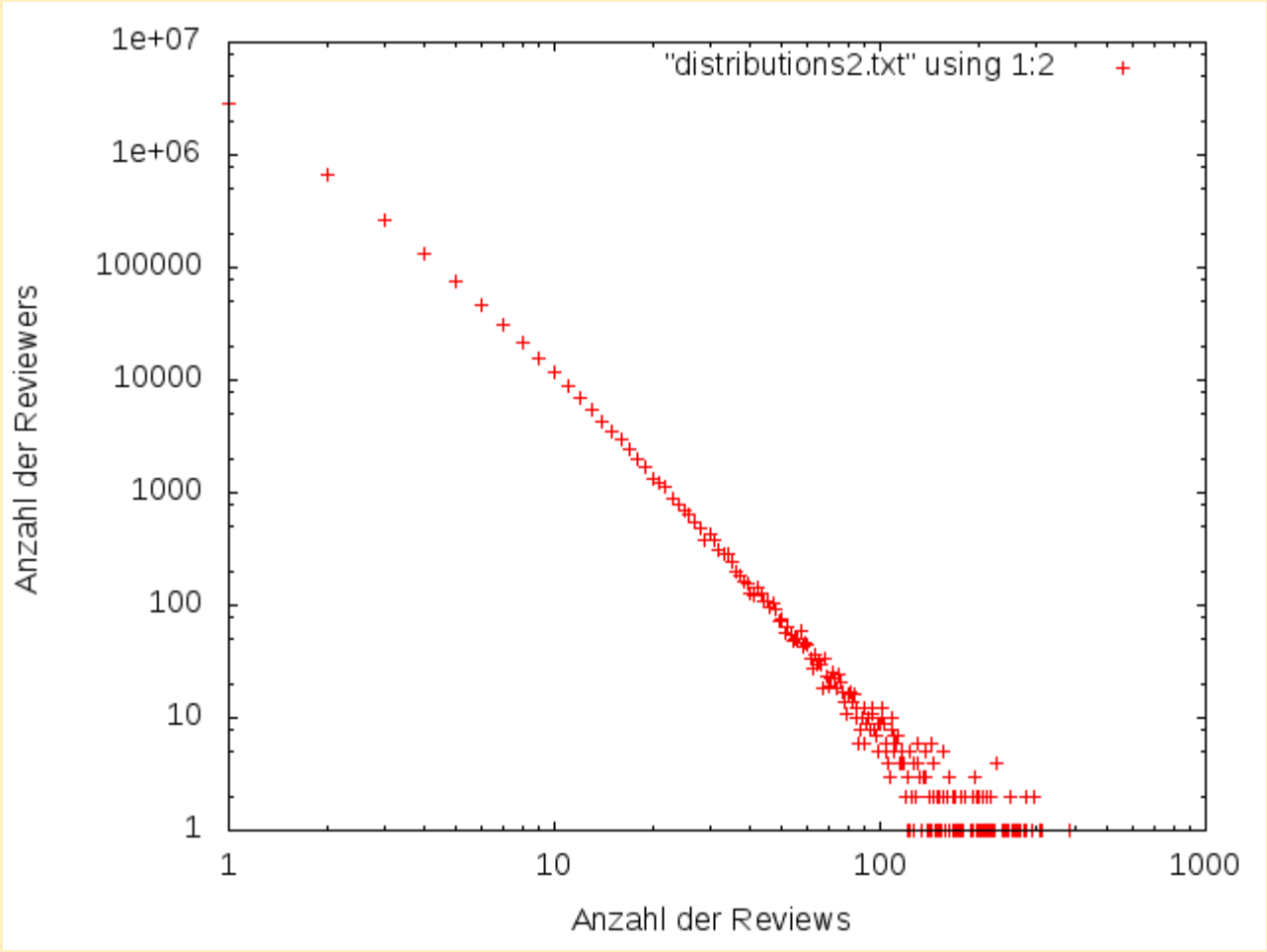

Figure: Reviewverteilung

Reviewerverhalten bei steigender Reviewanzahl

- Gruppierung der Nutzer nach Reviewanzahl beibehalten
- Merkmale: Bewertung, Wortanzahl, Zeichenanzahl, hilfreich
- Innerhalb einer Gruppe Merkmale arithmetisch mitteln

Reviewerverhalten bei steigender Reviewanzahl

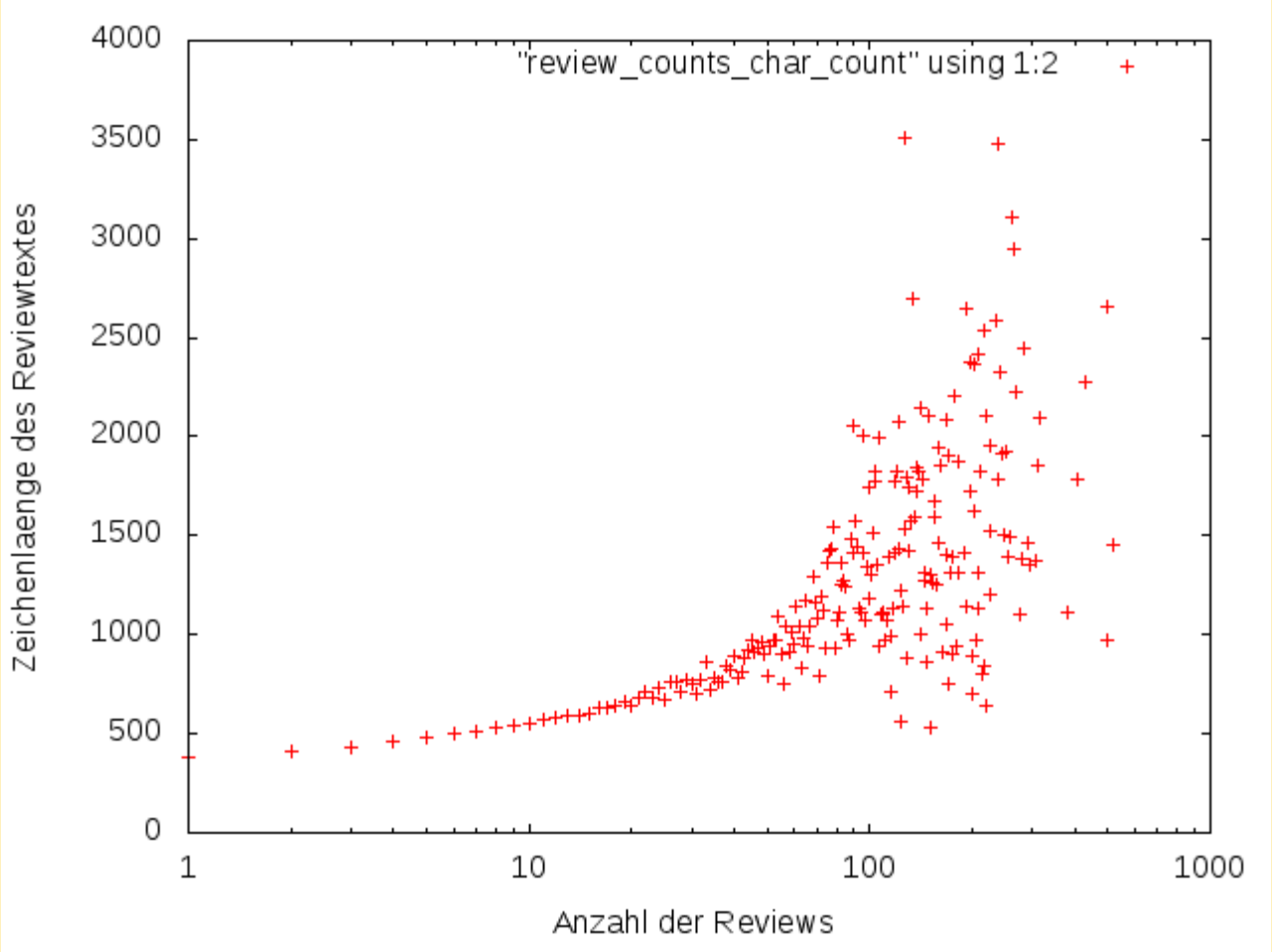

Figure: Reviewverteilung Reviewanzahl und Zeichenanzahl

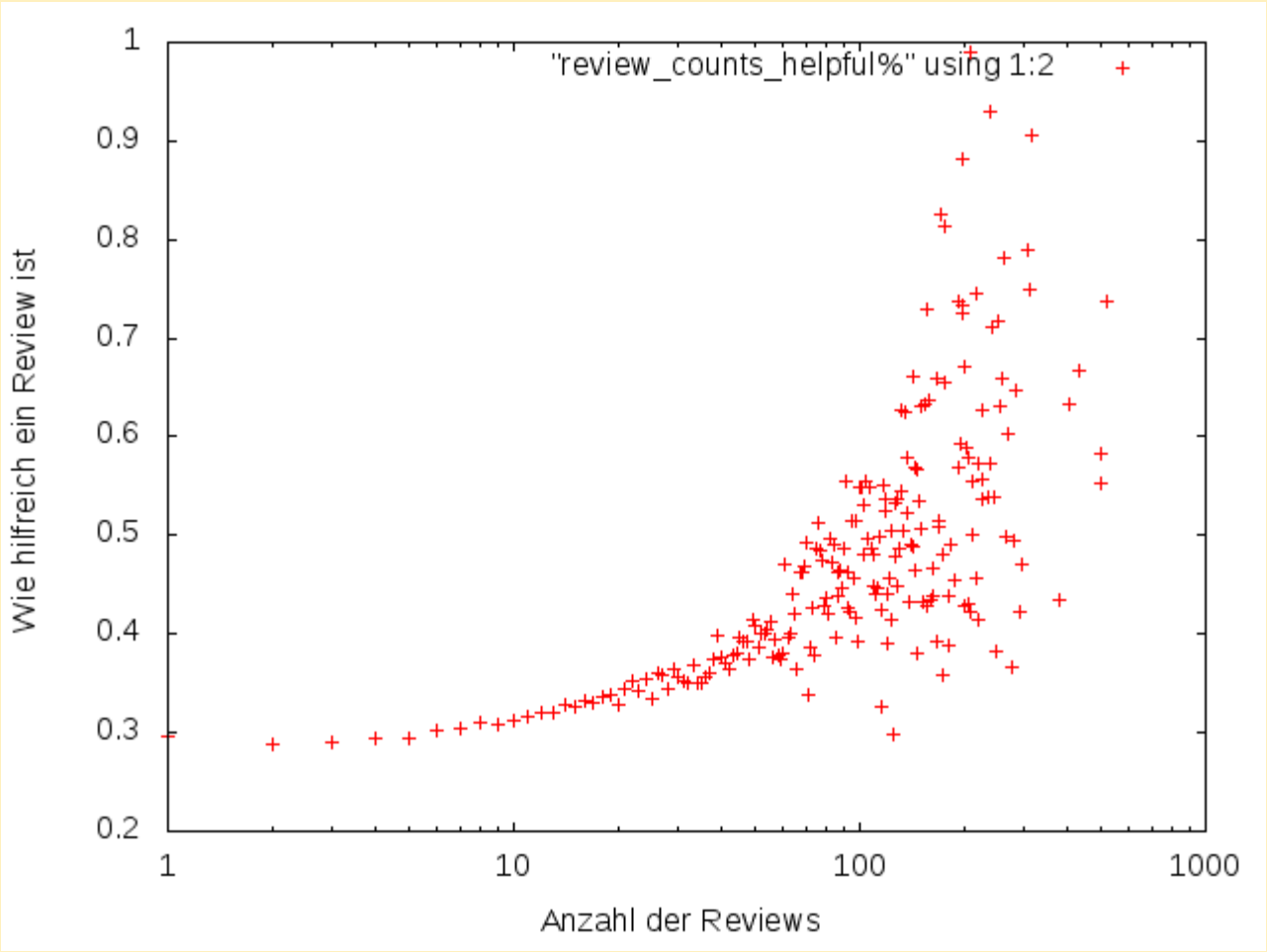

Figure: Reviewverteilung Reviewanzahl und Hilfreichbewertungen(%)

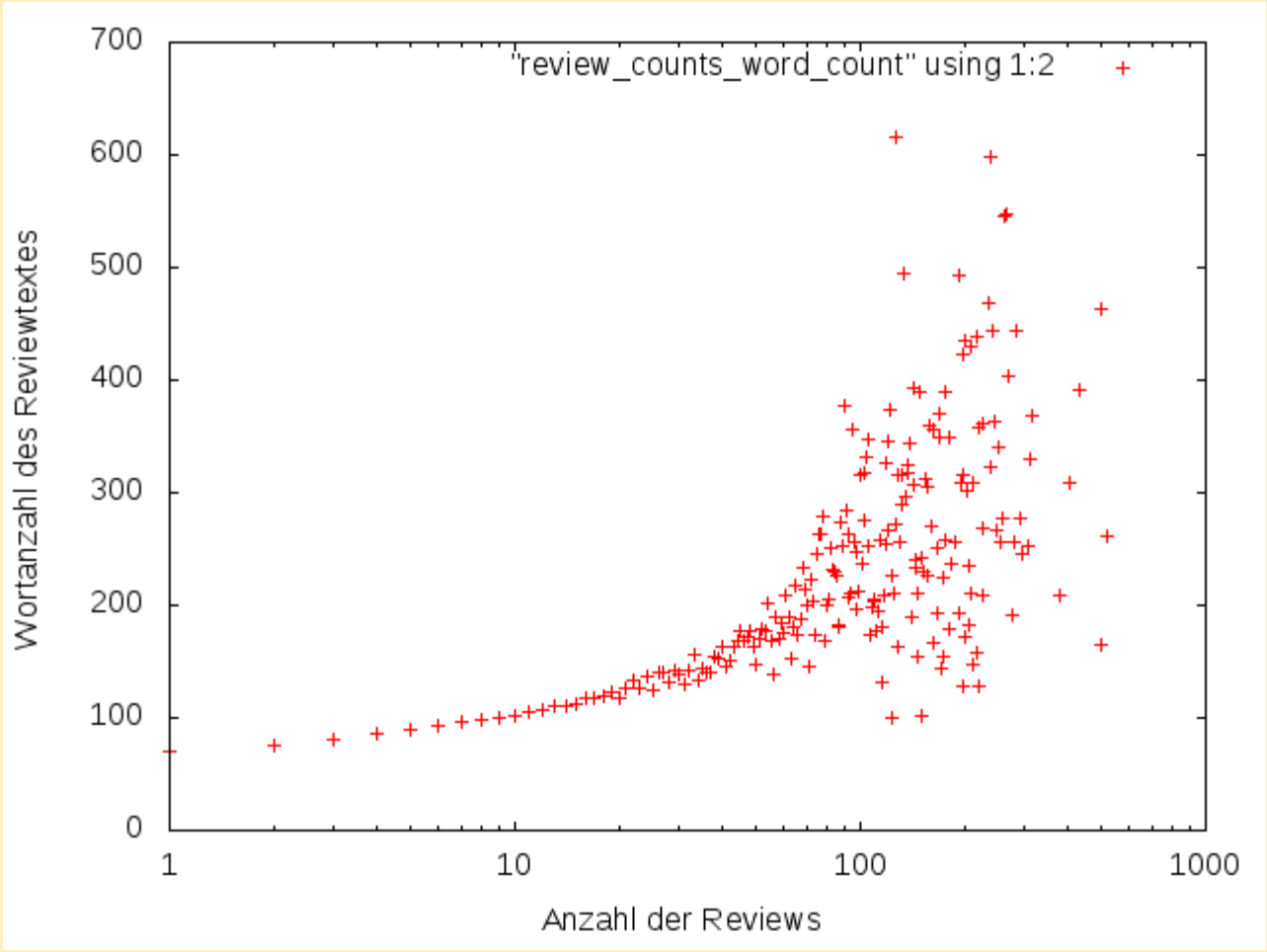

Figure: Reviewverteilung Reviewanzahl und Wortanzahl

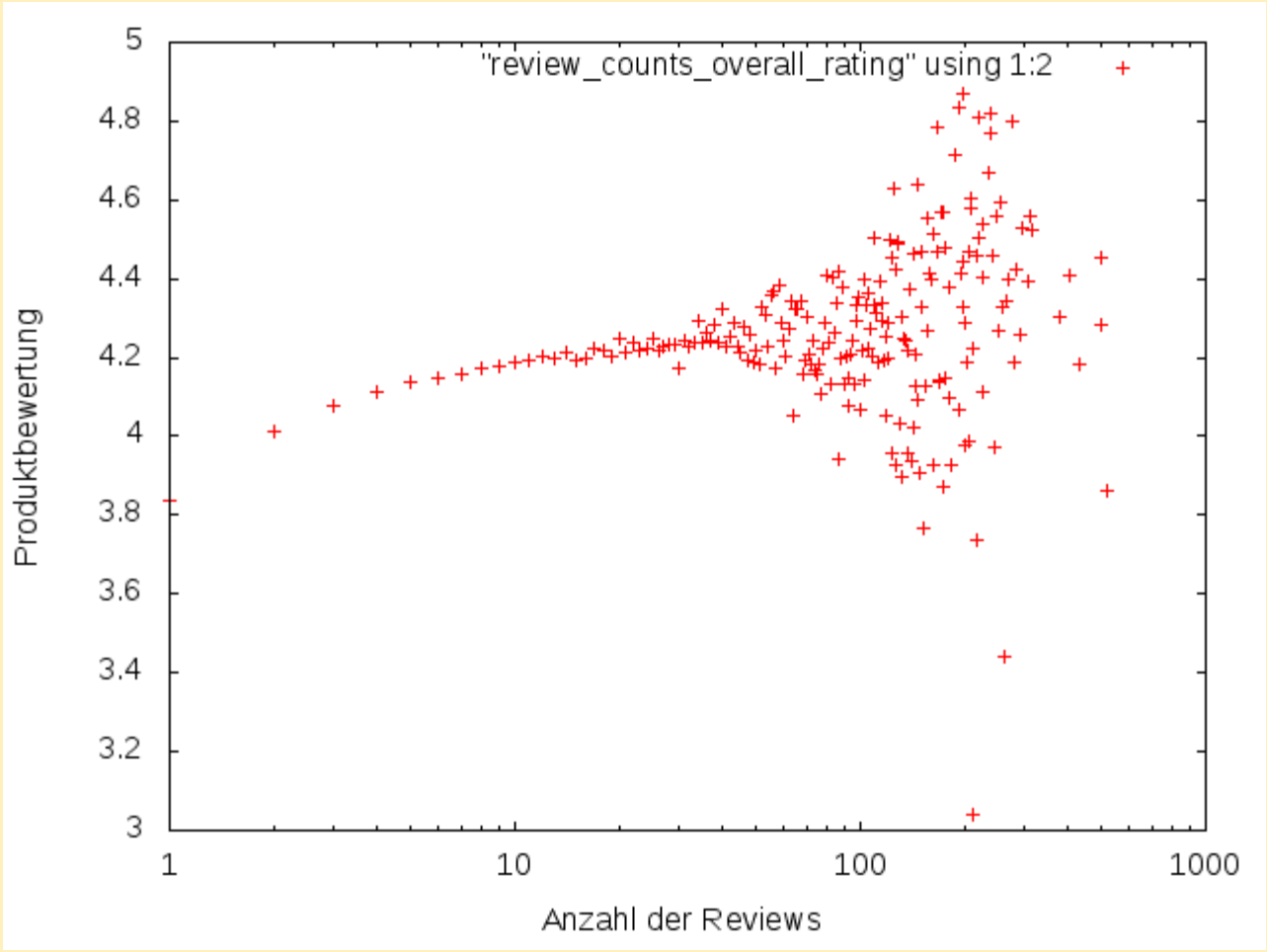

Figure: Reviewverteilung Reviewanzahl und Produktbewertung

Textlänge und Produktbewertung

- Analog über Reviewlänge gruppiert
- Produktbewertungen je Gruppe gemittelt
- Ein Bewertungstief bei ca 180 Wörtern
- Ab 1000 Wörtern ungenau

Textlänge und Produktbewertung

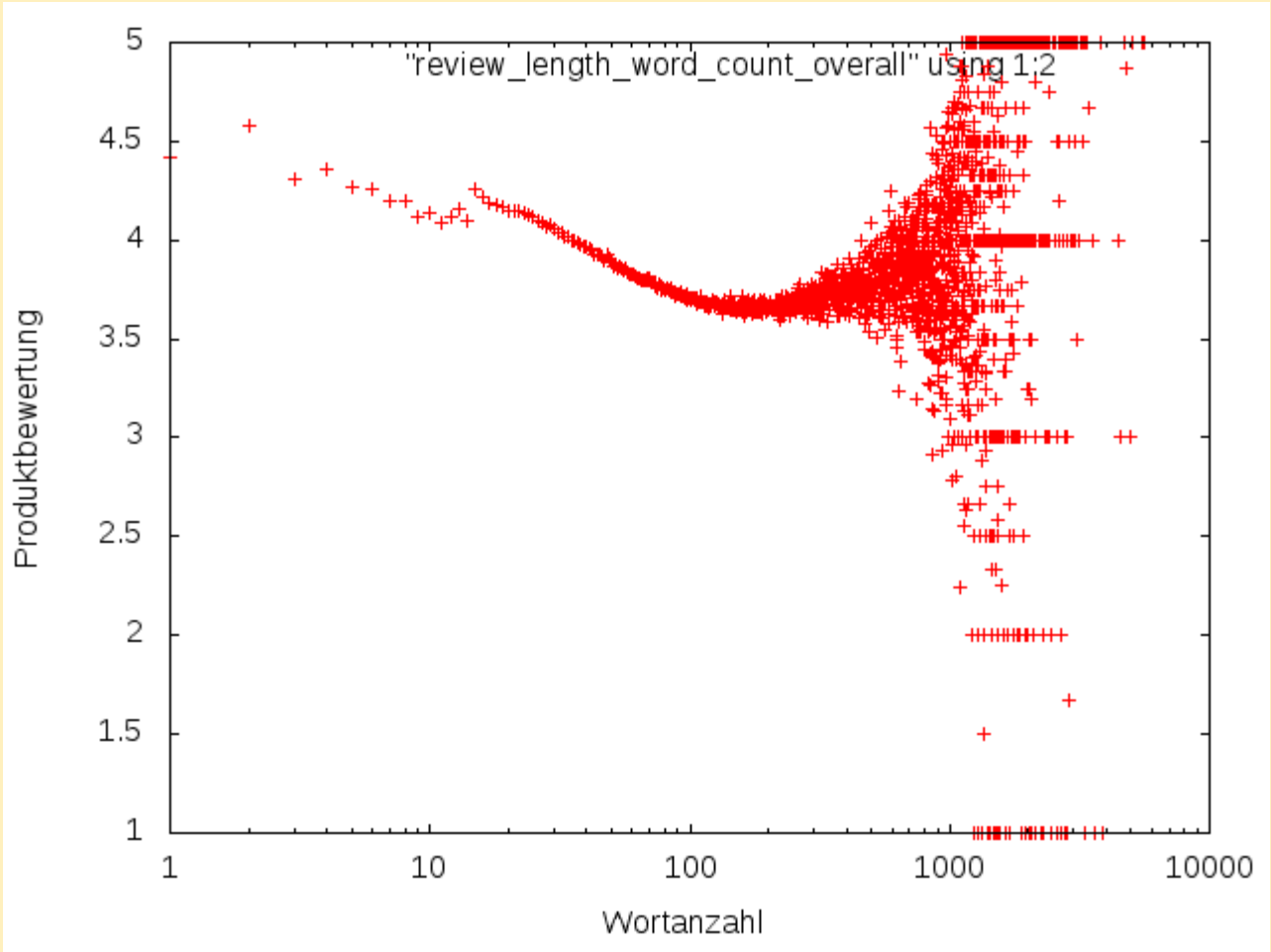

Figure: Reviewverteilung

Textlänge und Produktbewertung

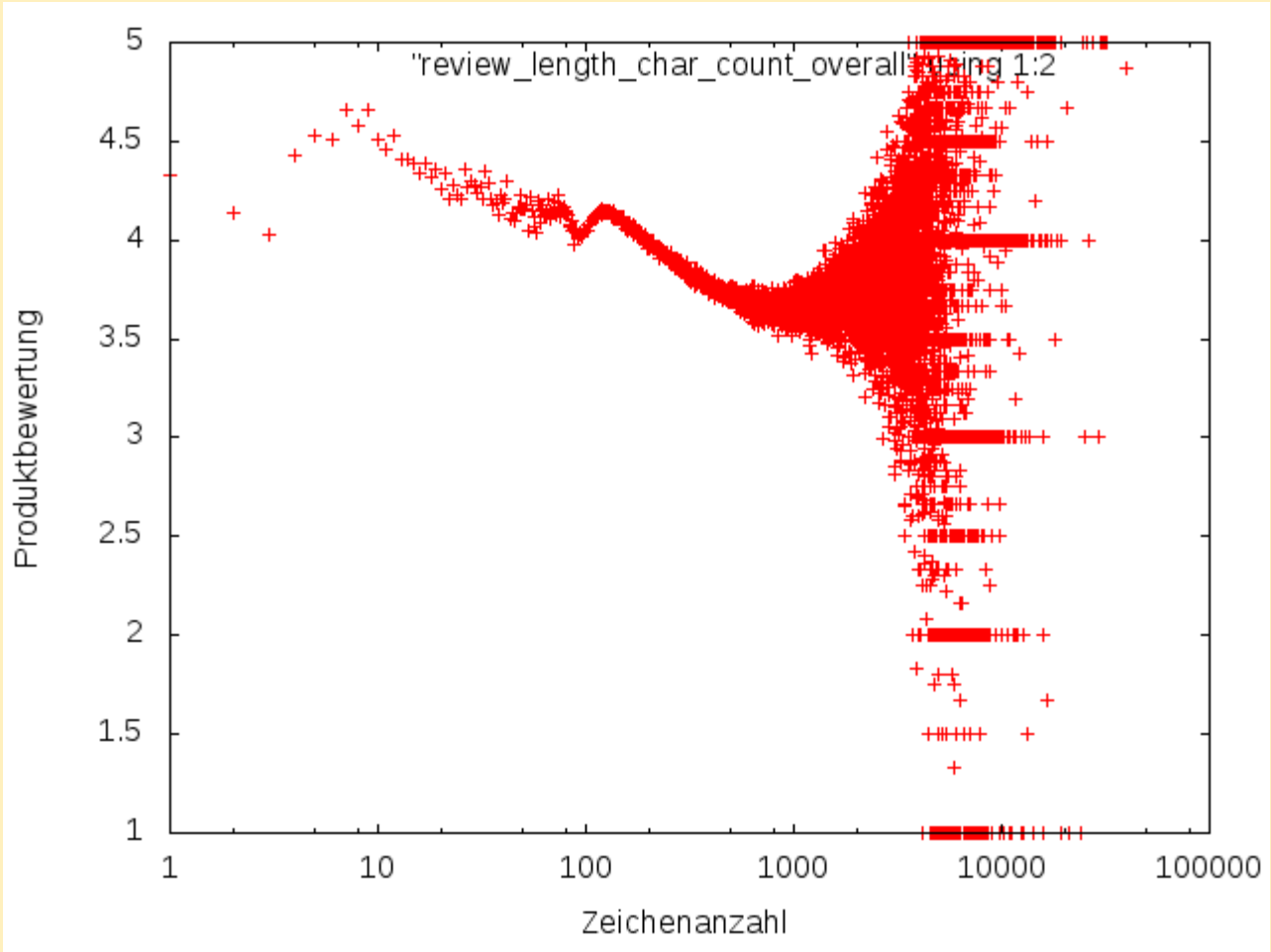

Figure: Reviewverteilung

Clustering

- Nutzer mit seinen Reviews erhält Merkmalsvektor
- Merkmale:
  - Wortanzahl des Bewertungstextes
  - Zeichenanzahl des Bewertungstextes
  - Zeitstempel des Reviews
  - Zeichenanzahl des Reviewtitels
  - wie hilfreich das Review war prozentual
  - wieviele das Review ingesamt nützlich fanden
  - Produktbewertung
  - Reviewanzahl
- Clusterverfahren mit Distanzmaß zwischen allen Reviewern $\rightarrow$ verworfen, da zu aufwändig
- PCA
- K-Means mit $k = 2, 44$ und 65 gewählt

Danke