seit 1558

Big Data

Projekt: Amazonreviews

Matthias Körschens & Kevin Reinke

Friedrich-Schiller-Universität Jena

January 30, 2017

Ziel

Explorative Datenanalyse von AmazonReviews. Im speziellen die Verteilung der Reviews, das Bewertungsverhalten der Nutzer sowie Clustern von Nutzern.

- 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: JASON

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 ihre eigenen Reviews zugeordnet
- Die meisten Nutzer bewerten wenig. 2,88 von 4,2 Millionen Nutzern die nur ein einziges Review abgaben.
- Zusammenhang exponentiell (Zipfsche Gesetzt)

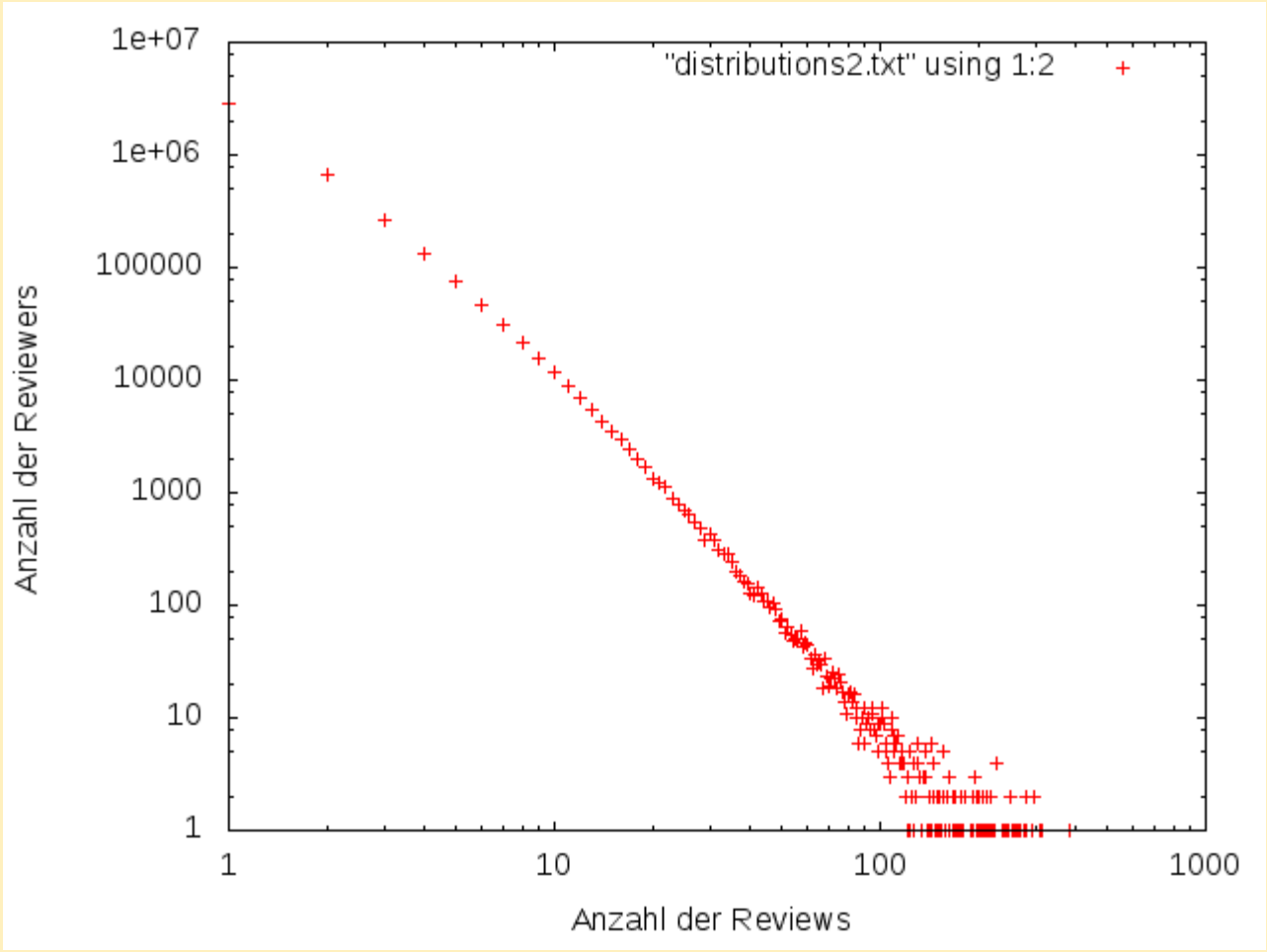

Figure: Reviewverteilung

Reviewerverhalten bei steigener Reviewanzahl

- Gruppierung der Nutzer nach Reviewanzahl beibehalten
- Merkmale: Bewertung, Wortzahl, Zeichenzahl, Hilfreich
- innerhalb einer Gruppe Merkmale arithmetisch mitteln

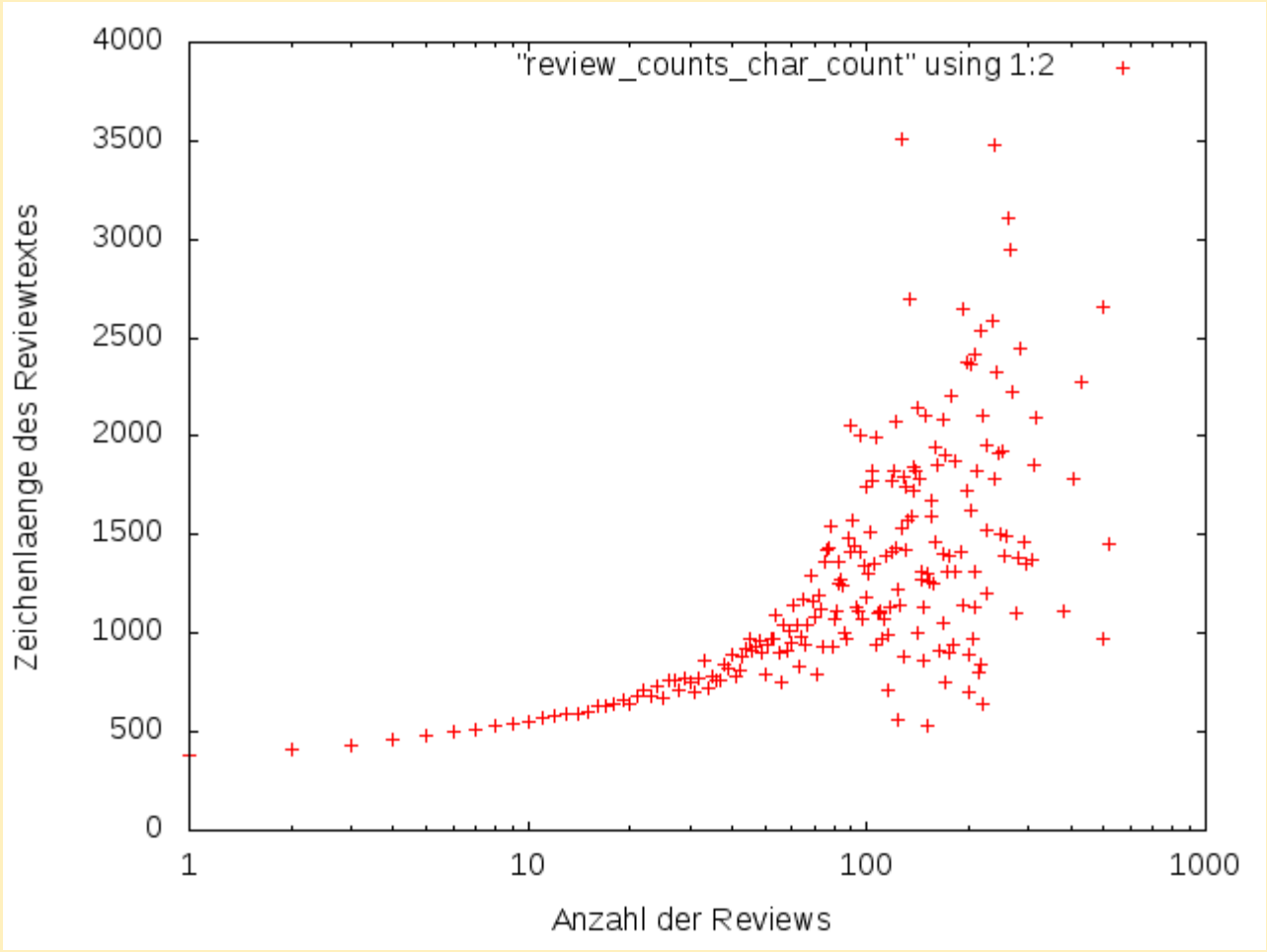

Figure: Reviewverteilung

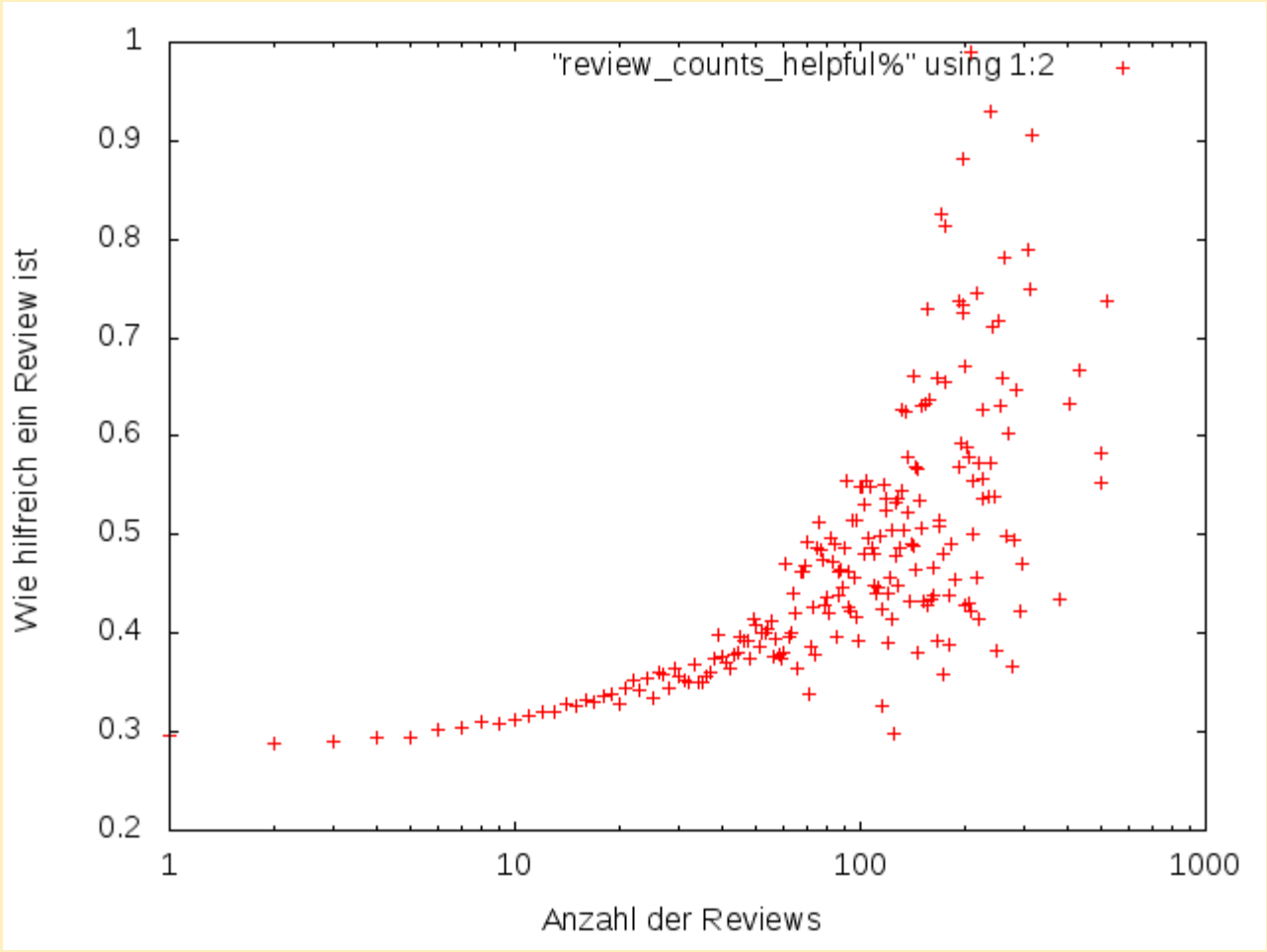

Figure: Reviewverteilung

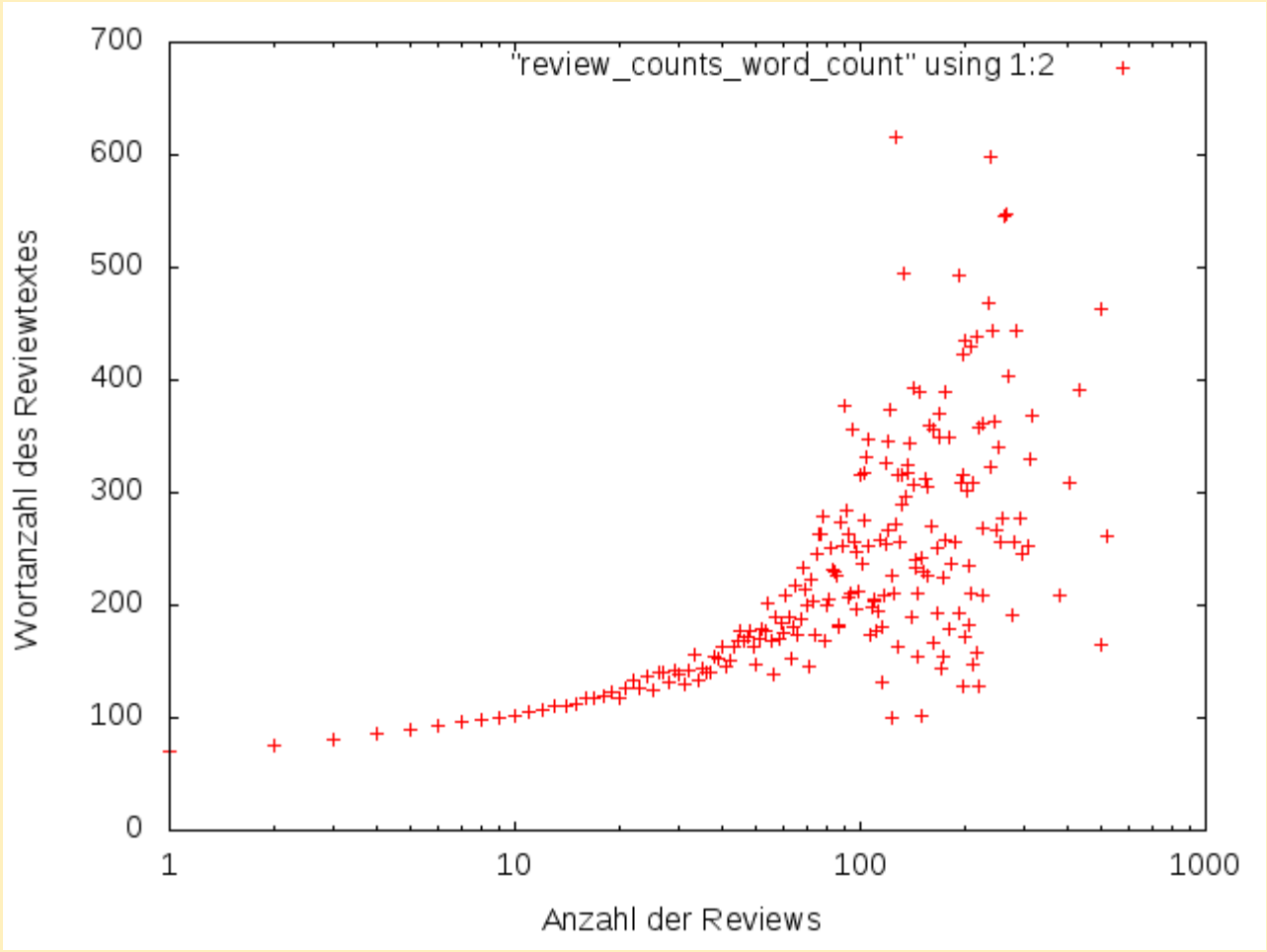

Figure: Reviewverteilung

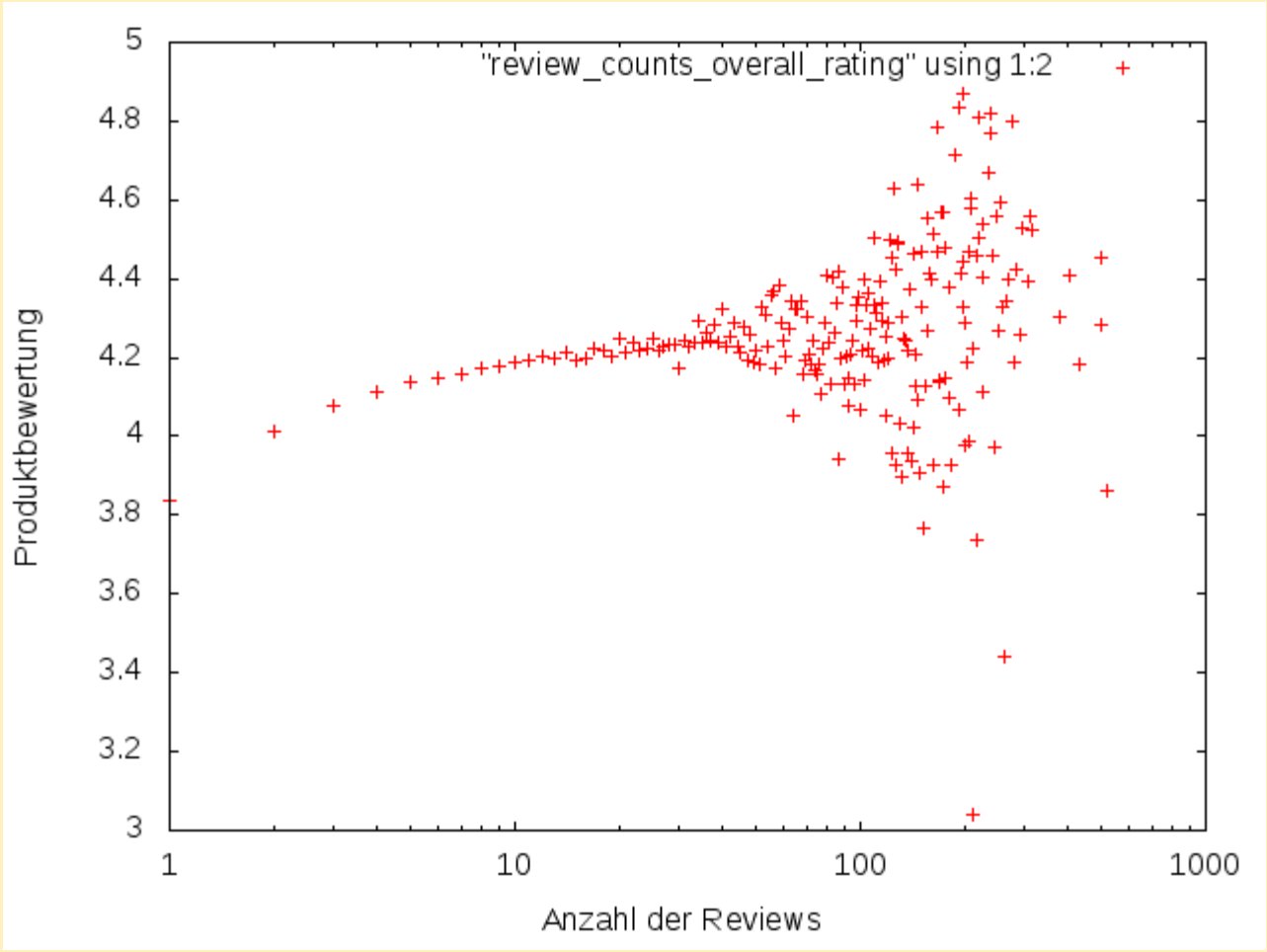

Figure: Reviewverteilung

Textlänge und Produktbewertung

- Analog über Textlänge Gruppiert
- Produktbewertungen je Gruppe gemittelt
- ein Bewertungstief bei ca 180 Wörtern
- ab 1000 Wörtern ungenau

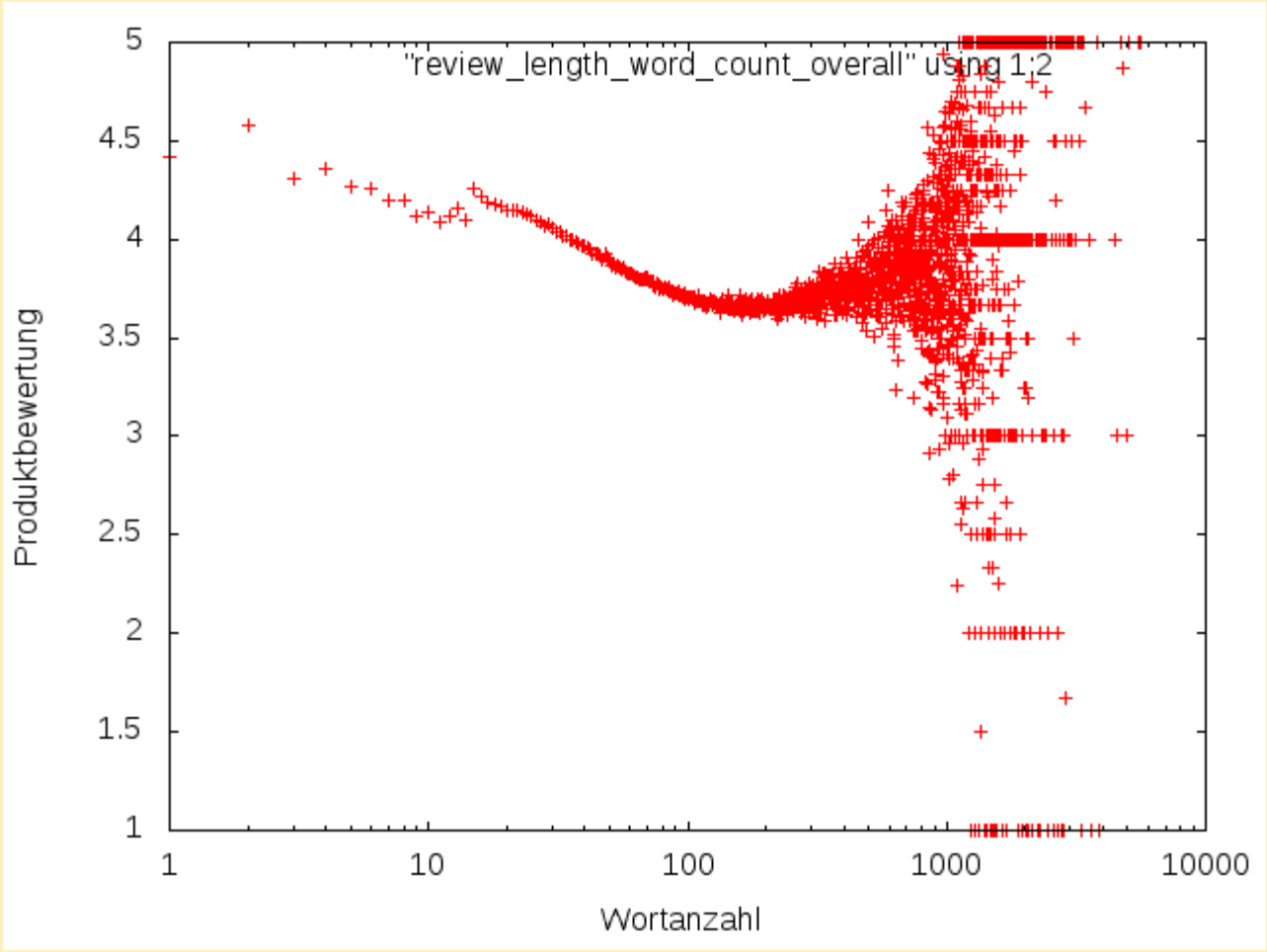

Figure: Reviewverteilung

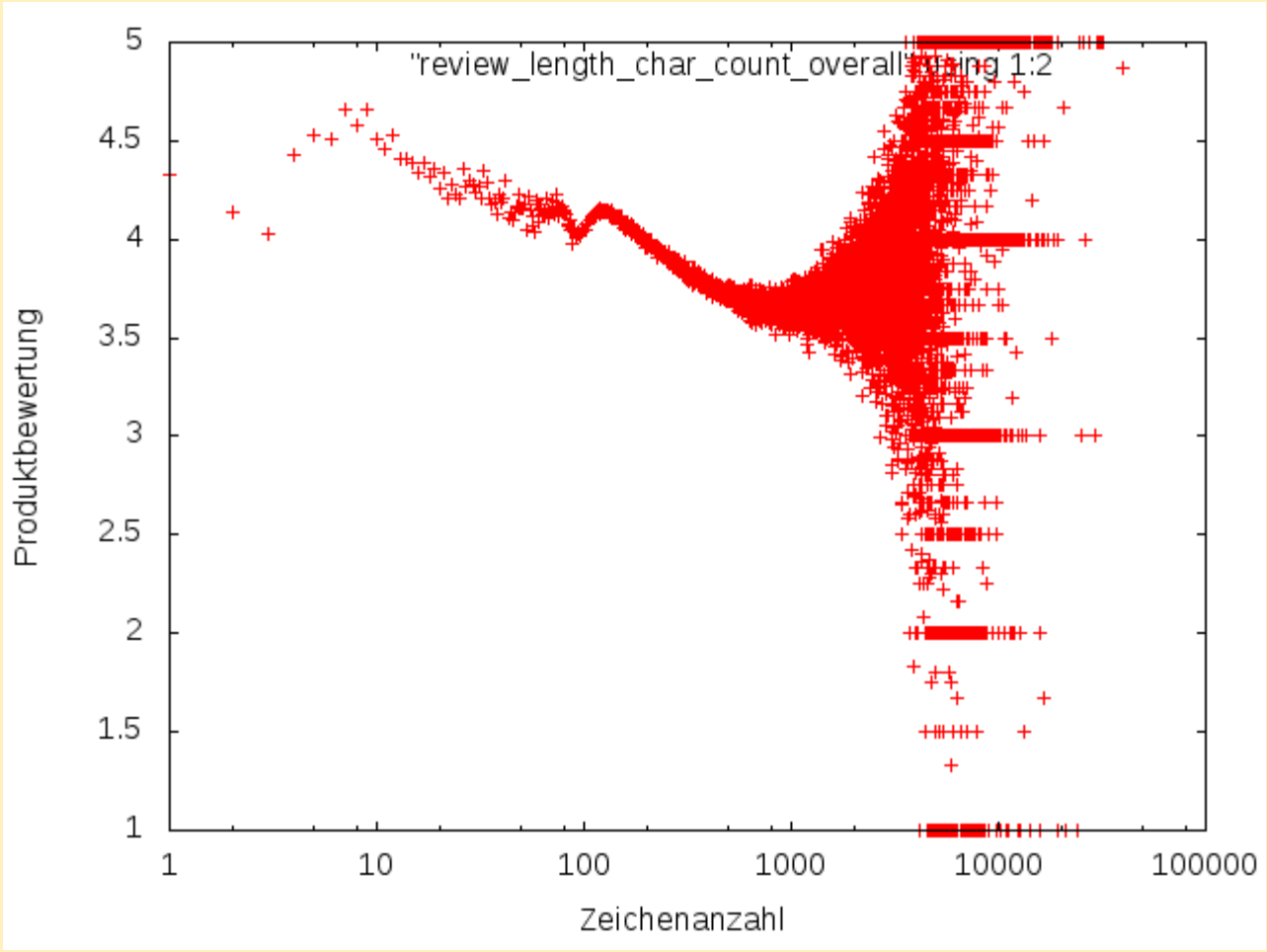

Figure: Reviewverteilung

- 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
- Clusterverfahren, mit Distanzmaß zwischen allen Reviewern$\rightarrow$ verworfen
- PCA
- K-Means mit $k = 2, 44$ und 65 gewählt

Danke