Wortschatz Zeitgeist

Wolfgang Otto, Thomas Döring, Max Kießling

Seminar Anwendungen der linguistischen Informatik

16. Juni 2015

Motivation

Algorithmen

Vergleich und Auswertung

Motivation

Wortschatzprojekt

- Bla
- Foobar
- Batz

Definition

Definition

Example

Beispiel

Algorithmen

Relative Häufigkeit

Something about relative frequency

TF/IDF

Something about TF/IDF

Poison

Something about poison

Z-Score

Something about Z-score

Zeitreihenanalyse

Definition (Zeitreihenanalyse)

Unter einer Zeitreihe versteht man die Entwicklung einer bestimmten Größe, deren Werte im Zeitablauf zu bestimmten Zeitpunkten oder für bestimmte Zeitintervalle erfasst und dargestellt werden

Maß: gleitender Mittelwert

- Glättet Zeit oder Datenreihen
- Erfolgt durch glätten hoher Frequenzanteile
- Es gibt ein Raster der größe n
- Es werden n Tage zusammenaddiert und dann durch n geteilt

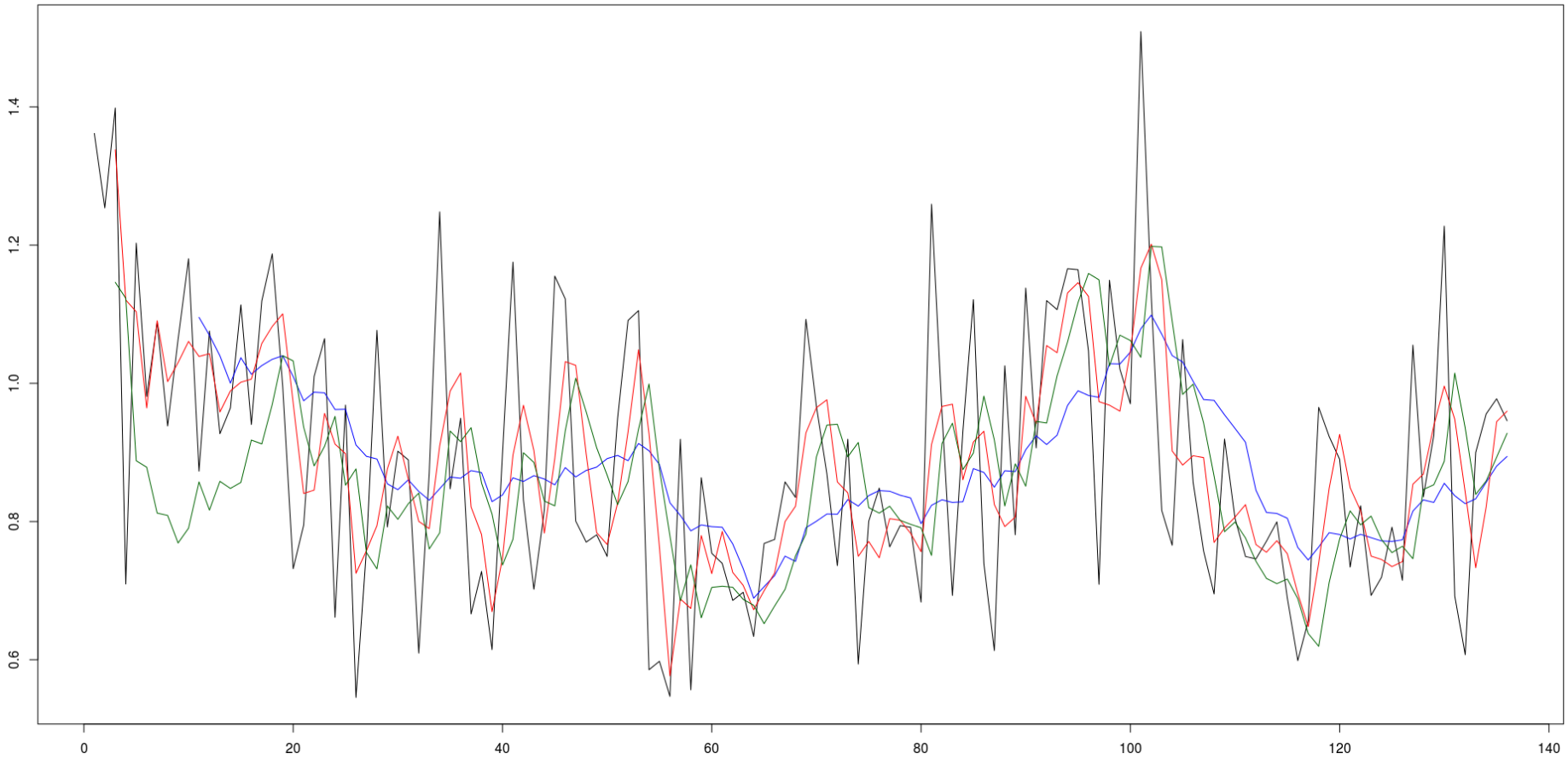
Wie hilft uns das weiter?

- Tritt ein Wort häufiger als sein Durchschnittswert an dem Tag auf kann das interessant sein.

Erster Ansatz: R

- Der erste Ansatz war ein R Programm welches den gleitenden Mittelwert ausrechnen sollte
- Problem: R verarbeitet Wörter einzeln
- 3 Mio. Wörter $\rightarrow$ 3 Mio. Transaktionen = MySQL Overkill
- Ausführungszeit würde mehrere Tage beanspruchen

Beispiel



Zweiter Ansatz: MySQL

- Der Zweite Ansatz ist es direkt in MySQL zu berechnen
- Problem: Inner Join auf selbe Tabelle (ca. 20 Mio Zeilen)
- Jeder Eintrag muss geprüft werden ob die Join Tabelle den Eintrag in der Größe des Rasters hat
- Eine Datums Differenz Tabelle kann das ganze jedoch beschleunigen

Finaler Ansatz: R BigTable

- Diesmal reshape der Tabelle
- Spalten = Wörter, Zeilen = Datumfelder, Wert = freq
- Darüber kann man das effizent einzeln berechnen
- Danach überführung in alte Strucktur und Speicherung

Vergleich und Auswertung

Zusammenfassung

Quellen (1)