Projekt Big Data

Abschlusspräsentation

Analyse von News-Artikeln – Sentimentanalyse mit Word2vec

Raffael Diestel, Maike Schubert

Inhalt

- Einleitung
- Ziele
- Realisierung
- Vorläufige Ergebnisse
- TODO

Einleitung

- Was ist eine Sentimentanalyse?
- Was ist Word2Vec?

Was ist eine Sentimentanalyse?

- Teilgebiet des Text Mining
- Analyse/Erkennung der Stimmung eines Textes
- Ziel: geäußerte Haltung als positiv, negativ oder auch neutral zu erkennen
- Methoden aus Statistik, maschinellem Lernen und Natural Language Processing

Was ist Word2vec?

- Modell zur Vektordarstellung von Wörtern → „word embeddings“
  - Input: großer Textkorpus
  - Konstruktion eines Vektorraums
  - Abbildung von Wörtern/Wortgruppen auf Vektoren reeller Zahlen
  - Wortvektoren so positioniert, dass Wörter mit gemeinsamem Kontext im Textkorpus im Raum nahe beieinanderliegen
  - Annahme: Wörter, die im selben Kontext auftreten, haben gemeinsame semantische Bedeutung

Ziele

Ziele des Projekts

herausfinden:

- Wie funktioniert eine Sentimentanalyse mit Word2vec?
- Inwiefern ist eine Sentimentanalyse (mit Word2vec) bei News-Artikeln möglich?
- Was lässt sich erkennen (Sentiments?)
- Sind Unterschiede bei bestimmten Themen oder den unterschiedlichen News-Portalen erkennbar?

Realisierung

- RSS-Feeds
- Daten
- Extraktion des Textes
- Word2vec
- Sentiment-Analyse-Ansatz
- Grafik zum Vorgehen

RSS-Feeds

-Crawler gestartet mit 9 RSS-Feeds

-jeweils alle Artikel, nicht nach Themen gefiltert

FOCUS (www.focus.de) -http://rss.focus.de/fol/XML/rss_folnews.xml	ZEIT online (http://newsfeed.zeit.de) -http://newsfeed.zeit.de/index
Spiegel Online (www.spiegel.de) -http://www.spiegel.de/schlagzeilen/index.rss	ZDF (www.zdf.de) -https://www.zdf.de/rss/zdf/nachrichten
T-Online (http://www.t-online.de) -http://rss1.t-online.de/c/11/52/20/74/11522074.xml	FAZ (www.faz.net) -http://www.faz.net/rss/aktuell/
Tagesschau (www.tagesschau.de) -http://www.tagesschau.de/xml/atom/	Sueddeutsche (www.sueddeutsche.de) http://rss.sueddeutsche.de/app/service/rss/alles/index.rss?output=rss
WELT (www.welt.de) -https://www.welt.de/feeds/latest.rss	

Daten

- 108GB Website-Downloads
- 61.343 Artikel
- 15 Mio. Wörter
- SentiWS
  - 1650 positive/1818 negative Wörter Grundformen
  - Inklusive Flexionsformen insgesamt 15.649 positive/15.632 negative Wörter

Extraktion des Textes

- Mit Beautiful Soup 4
  - Python Library zum Extrahieren von Daten aus HTML- und XML-Dateien
  - Unterstützt den HTML-Parser aus Pythons Standard-Library
- Zugriff auf relevante HTML-Tags
- Bereinigen des Textes
- Speichern in Datei

- 108GB Website-Downloads → 158MB Reintext

Word2vec

- Text aller Artikel einer Seite laden
- Aufsplitten in Sätze und Stoppwörter entfernen (NLTK)
  → Liste aus Wortlisten
- Word2vec Modell erstellen und trainieren
  → Wortvektoren

Word2vec-Beispielanfragen

```
>>> w2v.model.wv.similarity('Kanzlerin', 'Merkel')
0.917639468366345

>>> w2v.model.wv.most_similar('GroKo')
('Koalition', 0.939128994941714)

>>> w2v.model.wv.most_similar_cosmul(['Kanzlerin', Präsident'], ['Putin'])
('Merkel', 0.963726758956909 2)
```

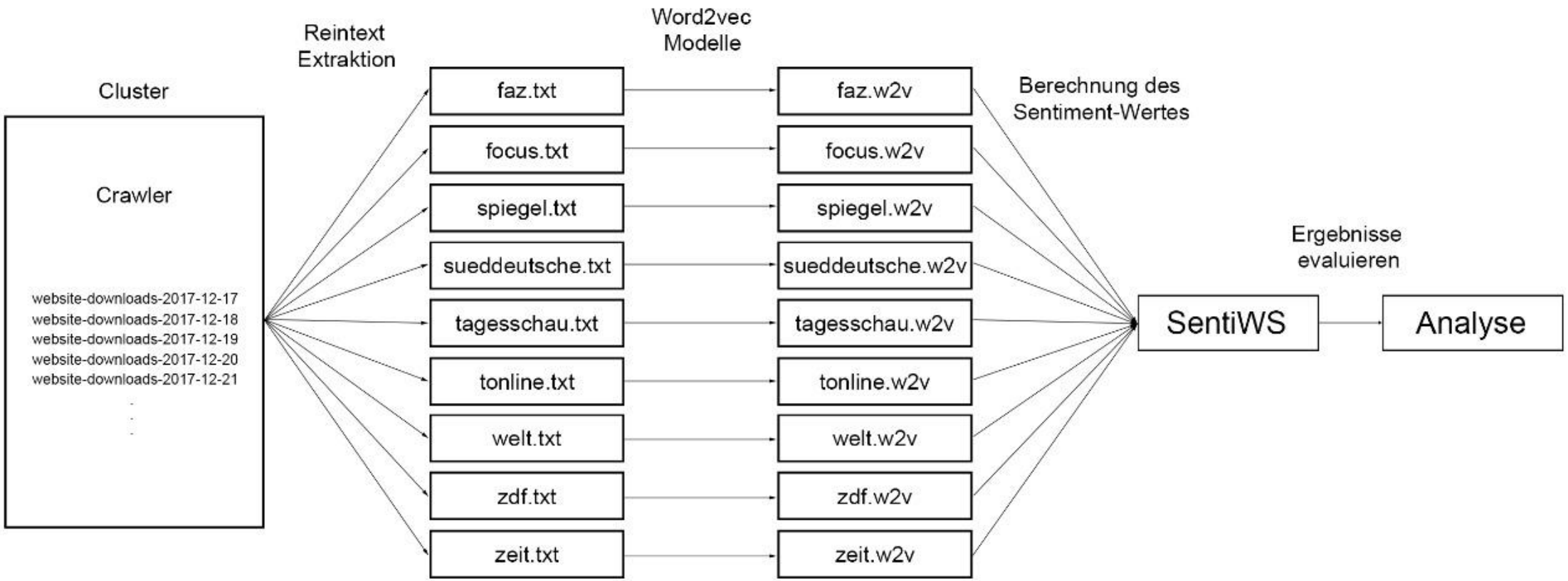
Sentiment-Analyse-Ansatz

- Verwendung der Ähnlichkeits-Funktion von Word2vec
- Berechnung eines Sentiment-Werts für jedes Wort im Artikel
- Sentiment-Wert basiert auf Ähnlichkeit zu den Wörtern aus SentiWS und deren Polarität

– Benötigt also gelabeltes Dataset (positiv, negativ)

– SentiWS → Wortschatz mit als positiv und negativ gelabelten Worten im Intervall [-1; 1]; wortschatz.informatik.uni-leipzig.de/de/download

Vorgehen grafisch dargestellt



Vorläufige Ergebnisse

- Ergebnisse
- Beispiele
- Probleme

Ergebnisse

- Hauptsächlich Einordnung als negativ
- Werbung positiv

- Durchschnittlicher Sentiment Score variiert stark zwischen den einzelnen Seiten (FAZ: -2300, Focus: -500)

- Größtenteils korrekte Einordnung in positive und negative Themen
  - →Eher Bewertung des Themas als des Sentiments
  - →Haltung des Autors kaum erkennbar
    - → evtl. Problem, dass News-Artikel zu sachlich geschrieben sind

Beispiel

- Durchschnittlicher Sentiment Score: 999,12 (positiv)
- Ausschlaggebende Wörter: hochwertige, interessante, besten, …

Donnerstag, 15.03.2018, 13:32

Hochwertige PCs und Tablets, interessante Spiele oder neue Kaffeevollautomaten - Das Angebot von Media Markt ist vielfältig. Entdecken Sie im aktuellen Media Markt-Katalog die Top-Deals der Woche.

Im aktuellen Media Markt-Prospekt finden Sie die besten Angebote der Woche. Media Markt bietet ein interessantes und vielfältiges Sortiment an innovativen Produkten, bei dem sowohl Anfänger als auch Technikspezialisten fündig werden. Nicht umsonst gehört Media Markt zu den erfolgreichsten Elektrofachmärkten in ganz Europa.

Quelle: focus.de

Beispiel

- Durchschnittlicher Sentiment Score: -1911,82 (negativ)
- Ausschlaggebende Wörter: Angreifer, getötet, Schüsse, …

Montag, 12.03.2018, 11:39

Bei einer Messerattacke vor der Residenz des iranischen Botschafters in Wien ist in der Nacht zum Montag der Angreifer getötet worden. Wie die Polizei mitteilte, griff ein 26-jähriger Österreicher einen Wachposten vor dem Gebäude mit einem Messer an.

Der Soldat habe zunächst vergeblich Pfefferspray eingesetzt und schließlich "mindestens vier" Schüsse aus seiner Dienstwaffe abgefeuert. Der Angreifer sei vor Ort gestorben.

Quelle: focus.de

Probleme

- Deutsche, gelabelte Daten benötigt
  - →Nicht viel zu finden, z.B. keine gelabelten Artikel
  - →SentiWS nicht optimal
- Kontext des Wortes wird nicht in die Bewertung mit einbezogen
  - →Wörter einzeln bewertet
- Gute Parameter für die Word2vec-Modelle finden
  - →Viel Trial & Error

TODO

- Berechnung des Sentiment Scores optimieren
  →Kontext betrachten, nicht nur einzelne Wörter
- Mehr/neue Parameter ausprobieren
- Projektbericht fertigstellen

Bildquellen

- https://www.focus.de/panorama/welt/wien-wachsoldat-erschiesst-messerangreifer-vor-irans-botschafterresidenz_id_8596180.html
- https://www.focus.de/shopping/prospekte/der-neue-media-markt-prospekt-die-top-angebote-der-kw-11-2018_id_7424451.html