Analysis of public textual business data to predict companies' condition

Max Lübbering
max.luebbering@tuhh.de

Supervised by:
Dr. Julian Kunkel, Dr. Patricio Farrell

October 15, 2017

Goals

- Build a huge data set containing broad range of business news
  - news articles
  - press releases
  - (agencies' ratings)
- Find features that describe a company's status
- Build model that predicts company's status from text data

Tools/Knowledge:

Data Collection:

- Python
- HTTP, RSS, XML
- Multi-threading

Data Analysis:

- Python, R, Jupyter
- Machine Learning
- Natural Language Processing
- Data Visualization

Methodology

- Crawler retrieves new articles via outlets' RSS feeds
- Preprocessing: Extracting text out of HTML downloads
- Develop models that
  - cluster articles having the same topic (e.g. all articles that have been published about Air Berlin's bankruptcy)
  - match articles to companies
  - estimate relevancy of news
    (e.g. shit storm on FB does not influence FB itself)
- Sentiment analysis of news articles
  - Use sentiment lexicon
  - Learn lexicon by stock value trends

Progress: Data Collection

Key figures of the data set:

- Started: 2017-07-10 17:57:21
- Stopped: 2017-10-02 01:59:42
- No. of outlets: 107
- No. of RSS feeds: 601
- Unique articles: 844,729
- Raw data size: 850GB
- Cleaned data size: 12GB

Progress: Data Collection

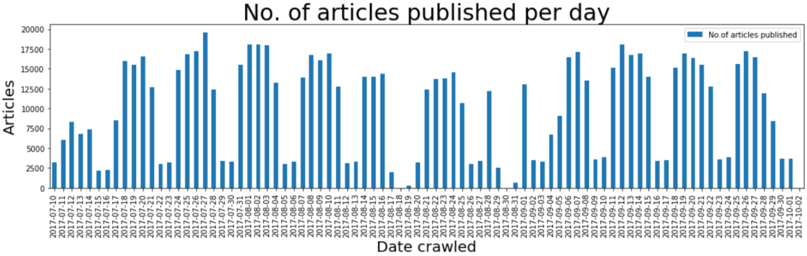

Figure: Number of articles published each day from 2017-07-10 to 2017-10-01