Chapter 3

Corpus

(German online news articles)

Data Pre-Processing

Document-Term Matrix

Bayrischer Innenminister will keine Altersgrenze mehr - »Verfassungsschutz soll Kinder beobachten - **Bild.de**

SPD: Bundestagswahl: Kandidat Schulz stellt Pläne zu Innerer Sicherheit vor - **FOCUS Online**

Linke-Parteitag in Hannover: Bedingt gesprächsbereit - **SPIEGEL ONLINE**

Wagenknecht sieht kaum Chancen für Rot-Rot-Grün – **stern.de**

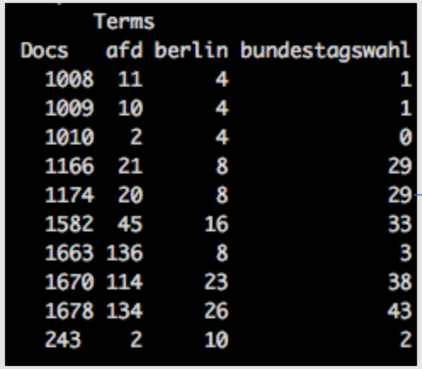
Klare Mehrheit: Bundestag will Einheitsdenkmal bis 2019 – **welt.de**

Anschlag in Kabul: Schulz will Abschiebungen nach Afghanistan aussetzen – **Zeit Online**

Includes the following steps:

1. Remove common words (Stopwords), punctuation, numbers and non-alphanumerical terms.
2. Stemming words to root words

The document term matrix is simply a mapping of how often each word appears in a particular article.



```
       Terms
Docs    afd berlin bundestagswahl
  1008   11      4              1
  1009   10      4              1
  1010    2      4              0
  1166   21      8             29
  1174   20      8             29
  1582   45     16             33
  1663  136      8              3
  1670  114     23             38
  1678  134     26             43
  243     2     10              2
```

Chapter 4

Generative Process

The algorithm analyzes the occurrences and attempts to identify the latent topics.

Generative Model: Latent Dirichlet allocation, where the prior distributions with globally shared mean parameters are replaced with means parameterized by a linear function of observed covariates.

Covariates: News Agency, Month

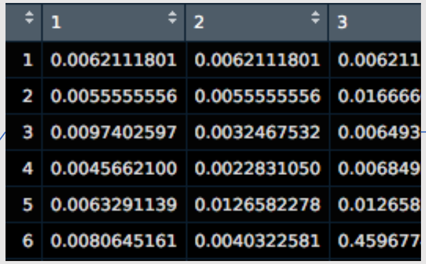
Algorithm: Gibbs Sampling

Chapter 5

Probabilities and Classification

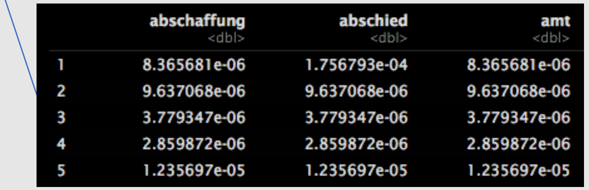
The output of the model is a set of probabilities mapping words to topics, and documents (news articles) to topics

Topic-document distribution θ



	1	2	3
1	0.0062111801	0.0062111801	0.006211
2	0.0055555556	0.0055555556	0.016666
3	0.0097402597	0.0032467532	0.006493
4	0.0045662100	0.0022831050	0.006849
5	0.0063291139	0.0126582278	0.012658
6	0.0080645161	0.0040322581	0.459677

Term-topic distribution ϕ



	abschaffung <dbl>	abschied <dbl>	amt <dbl>
1	8.365681e-06	1.756793e-04	8.365681e-06
2	9.637068e-06	9.637068e-06	9.637068e-06
3	3.779347e-06	3.779347e-06	3.779347e-06
4	2.859872e-06	2.859872e-06	2.859872e-06
5	1.235697e-05	1.235697e-05	1.235697e-05

Chapter 6

Estimation

We use the Topic-document distribution to estimate the conditional outcome distribution of Facebook shares v_i of document i on the topical prevalence θ_i of that document.

$$p[v_i|\theta_i]$$