scieneers
DRIVEN BY DATA

The political discourse on discrimination

Analysing 70 years of the Bundestag

scieneers
DRIVEN BY DATA

About me

Alexandra Wörner

Data scientist

Member in CorrelAid project team

Feminist

Diversity & inclusion

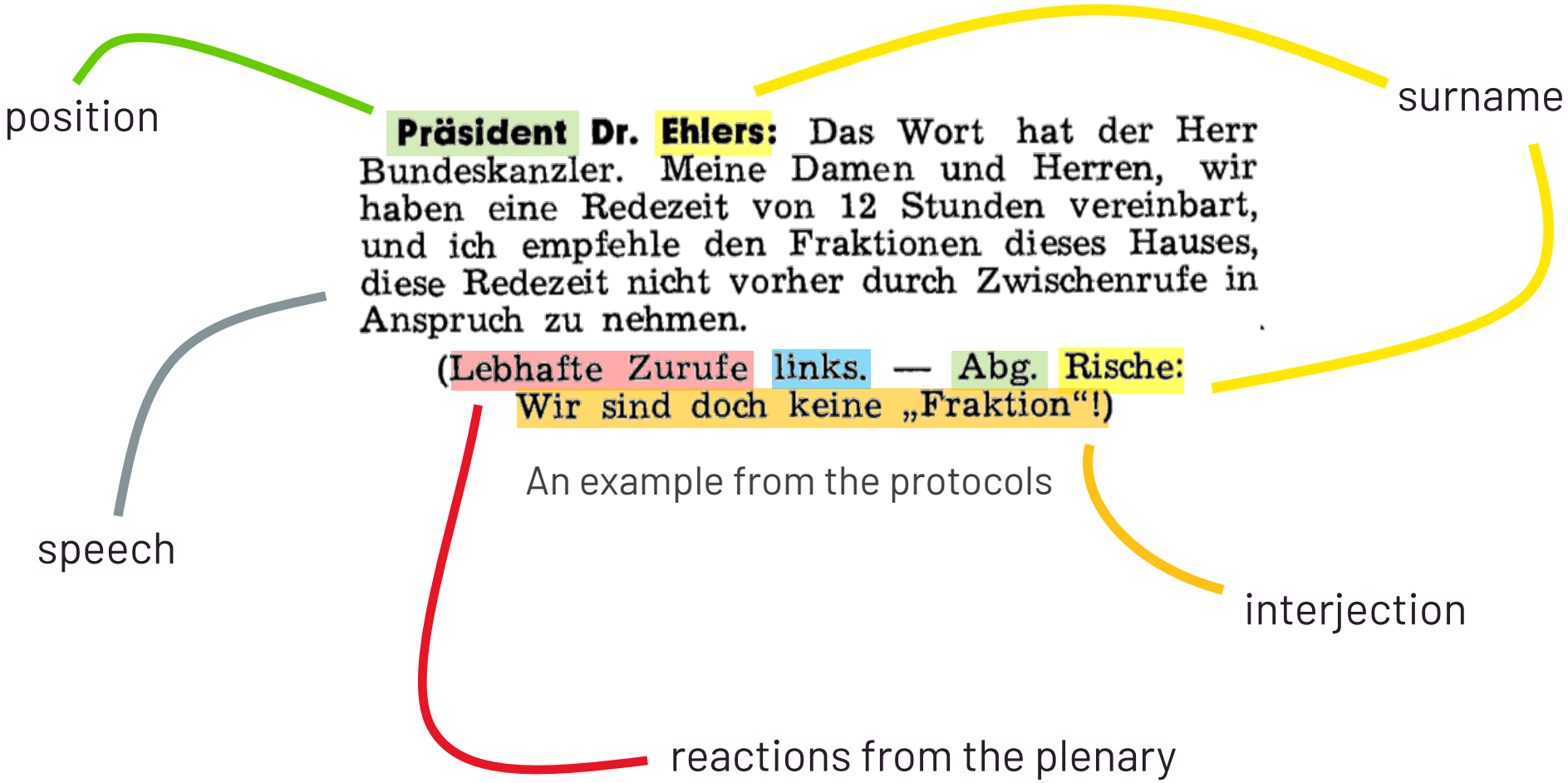
Interest in politics

What? facilitate machine readability of plenary protocols

How? techniques from computer science and computer linguistics

Why? German constitution states the publicness of the Bundestag

Elements of a speech extract



From the 190th session of the first German Bundestag
courtesy of Open Discourse (Florian Richter)

Facts & numbers

4265 protocols

4106 speakers

~ 900000 speeches

~ 2.1 Mio. interjections

~ 194 Mio. words

scieneers
DRIVEN BY DATA

What is fed into the models?

Google trains a multitude of models every day.

Tokenization & lemmatization

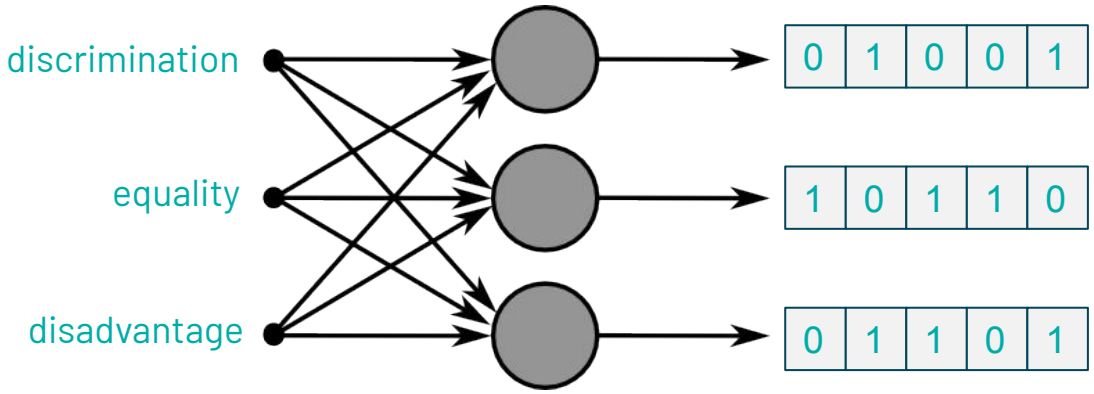
google - train - a - multitude - of - model - every - day - .

PROPN VERB DET NOUN ADP NOUN DET NOUN PUNCT

Model input

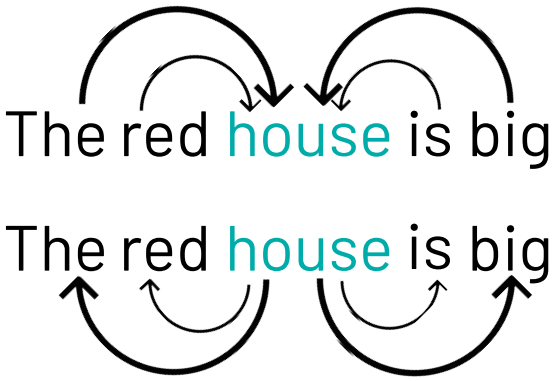
google - train - multitude - model - day

Semantic Analysis: Word2Vec

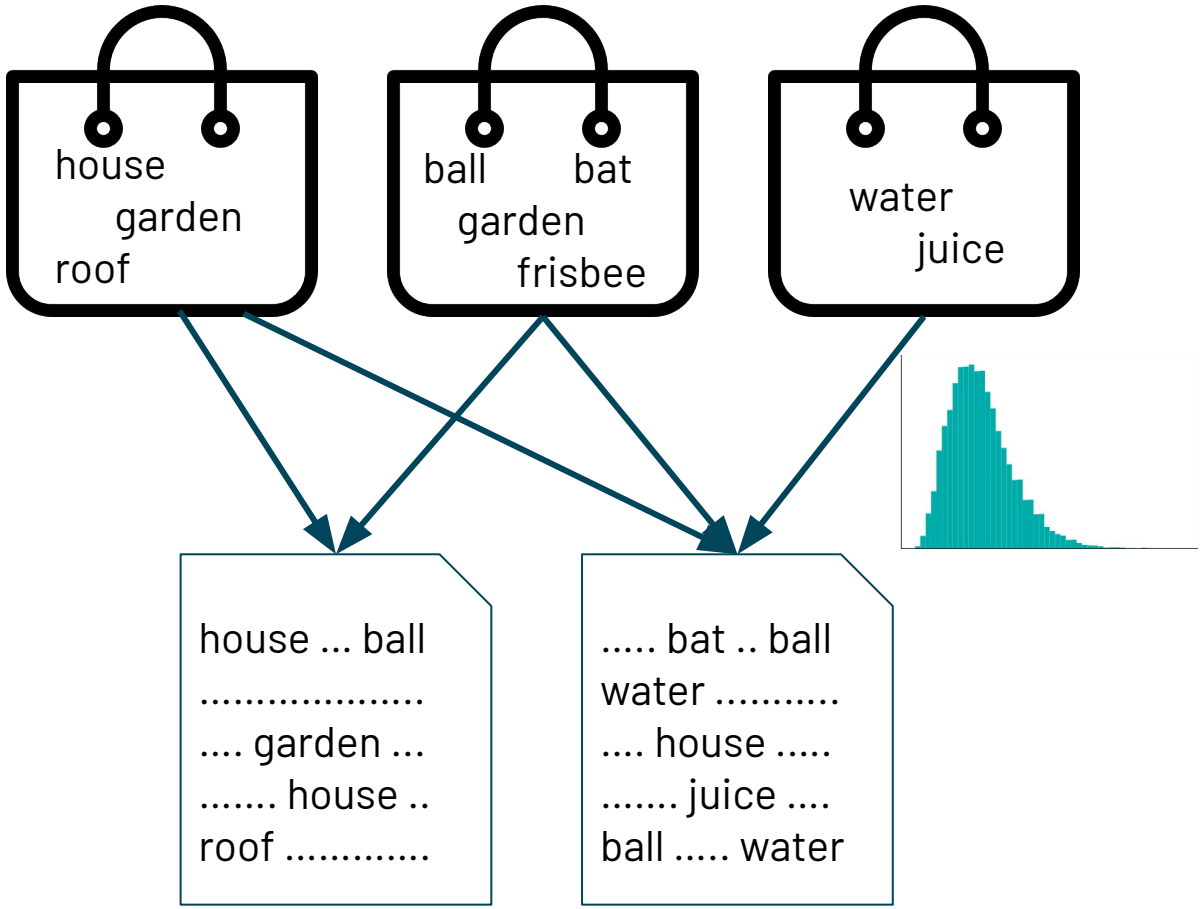


Two approaches

- Continuous bag of words
- Skipgrams



Semantic Analysis: Latent Dirichlet Allocation



	topic 1	topic 2	topic 3
house	0.8	0.1	0.1
garden	0.65	0.35	0
roof	0.95	0.02	0.03
ball	0.2	0.75	0.05
bat	0.15	0.82	0.03
frisbee	0.11	0.88	0.01
water	0.3	0	0.7
juice	0.06	0.04	0.9

scieneers
DRIVEN BY DATA

Demo time!

scieneers
DRIVEN BY DATA

Q & A

scieneers
DRIVEN BY DATA

Backup

scieneers
DRIVEN BY DATA

Implementation

python & jupyter

spacy for preprocessing

Word2Vec with gensim

scikit-learn contains LDA-implementation

plotly, wordcloud & pyLDAvis for visualisation