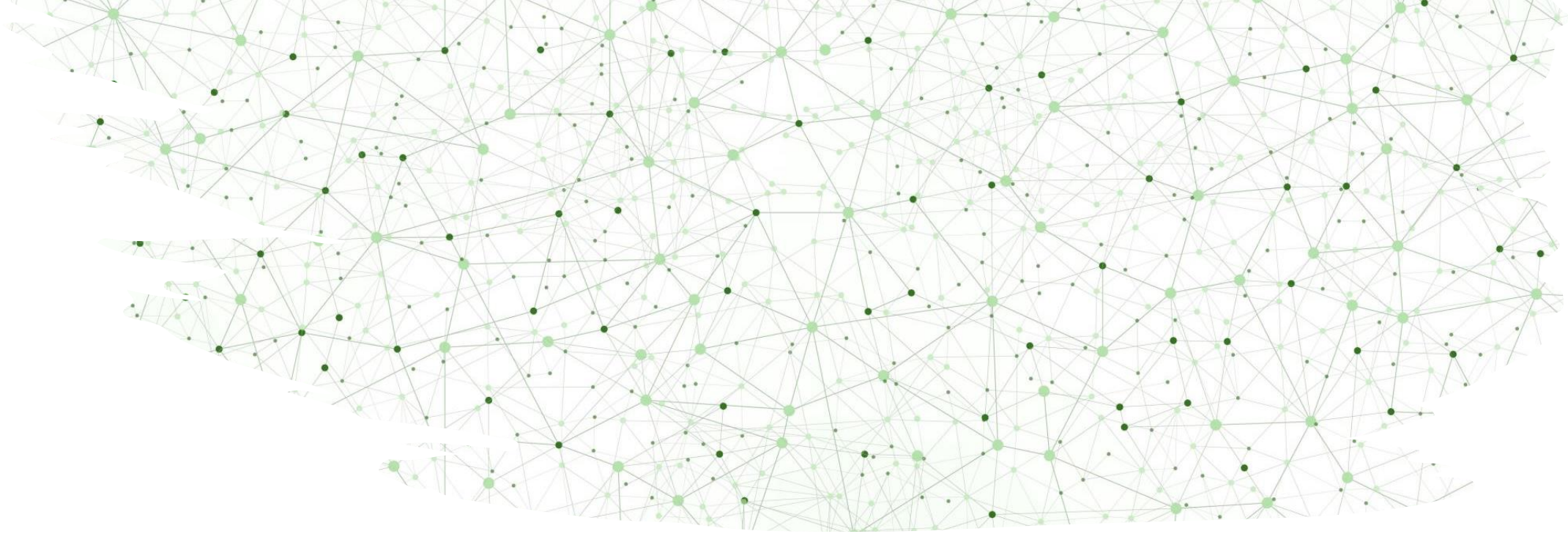
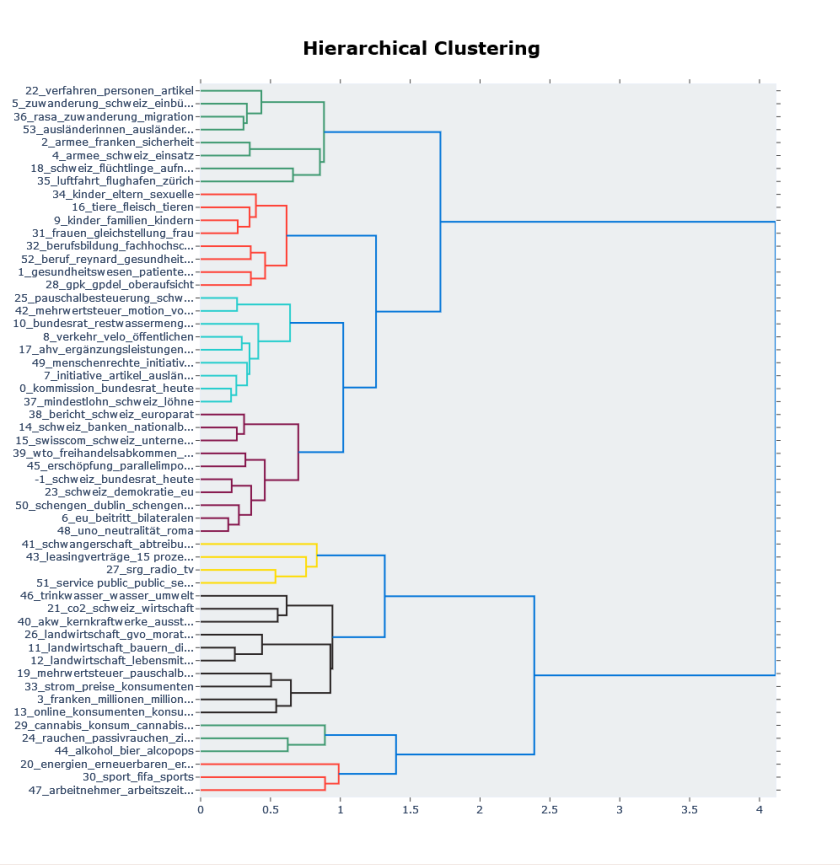
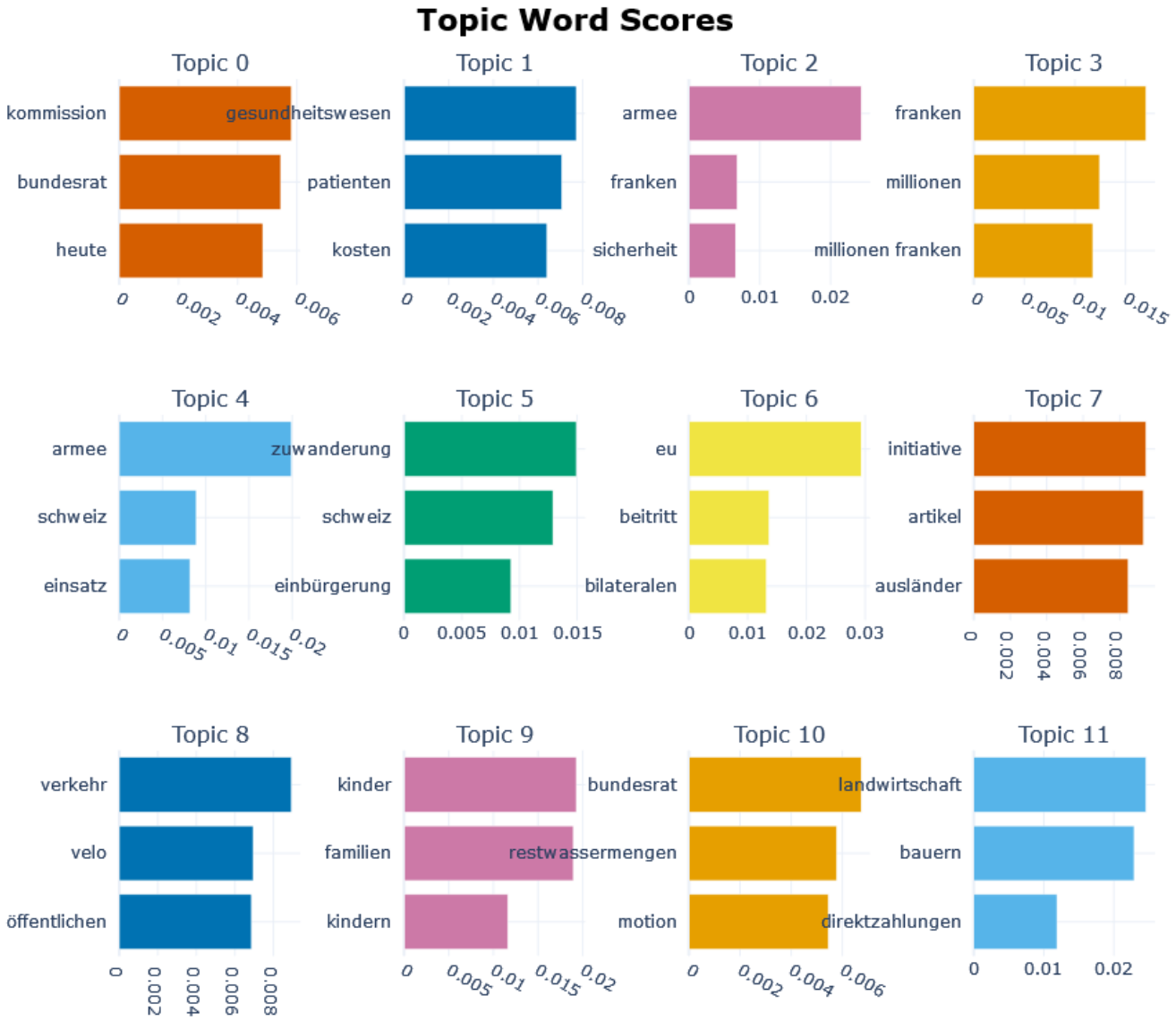


Report – Key Findings

TEXT MINING EXERCISE 3

Full Data Analysis on the speeches





Looking for similar topics for ‘co2’

SVP

```
topics,similarity = topic_model.find_topics("co2", top_n=5)
print(topics)
for top in topic:
  to = topic_model.get_topic(top) # lsva: (Leistungsabhängige Schwerverkehrsabgabe)
  print(to[0])
```

```
[16, 12, 9, 19, 21]
('co2', 0.0614439804064923)
('lsva', 0.01295256603120509)
('energien', 0.017631765368218717)
('luftfahrt', 0.028443119183008677)
('forschung', 0.018187095702797373)
```

SP

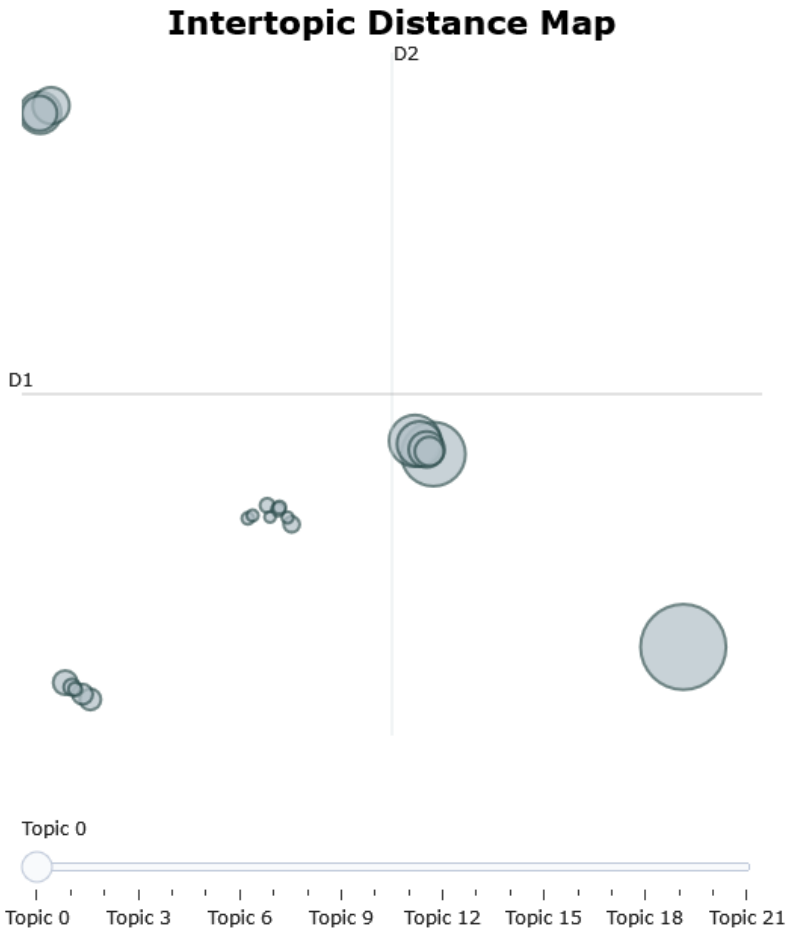
```
[187] topics,similarity = topic_model.find_topics("co2", top_n=5)
      print(topics)
      for top in topics:
        to = topic_model.get_topic(top)
        print(to[0])
```

```
[10, 26, 29, 16, 33]
('energien', 0.0206093647089728 82)
('erschöpfung', 0.04168836101149215)
('akw', 0.02703706663042125)
('gpk', 0.019364436399263007)
('natur', 0.01427017005899465)
```

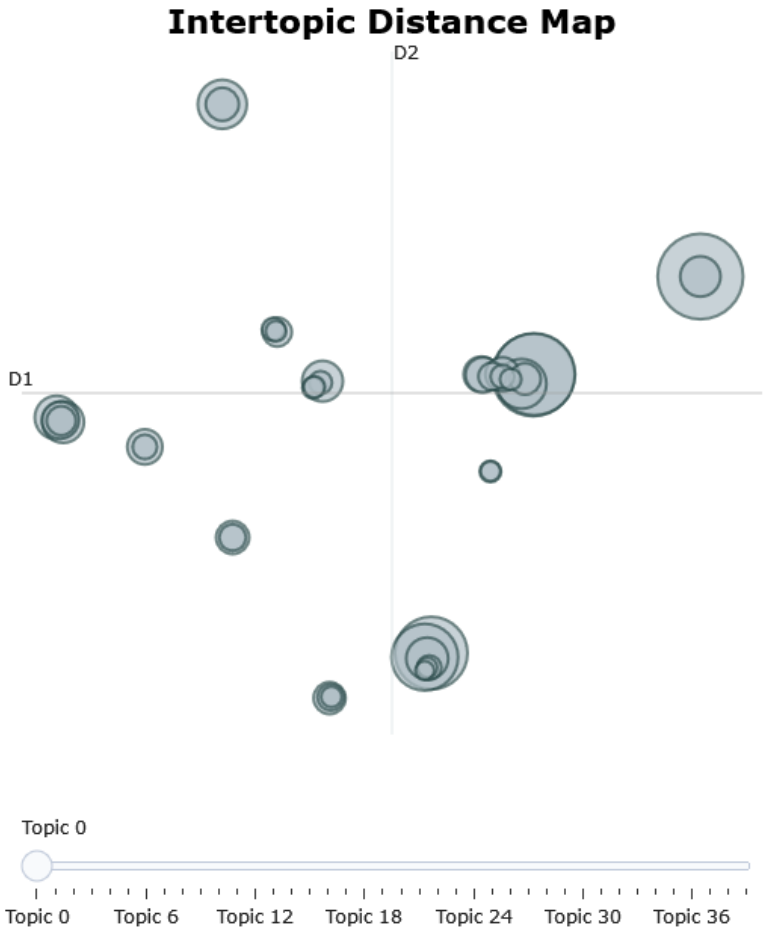
Interestingly, when we search for similar topics in the topic_model based on SVP-speeches vs. SP-speeches we clearly see a difference. It seems as if SVP ‘Co2’ related topics focus more on lsva(Scherverkehrsabgabe = shear traffic tax), luftfahrt (aviation) and forschung(research), which are topics more economic related. Whereas, in the SP-speeches similar topics to ‘CO2’ are very different and ‘Erschöpfung’= exhaustion in closely related to Co2.

Visualize topics, their sizes, and their corresponding words

SVP



SP



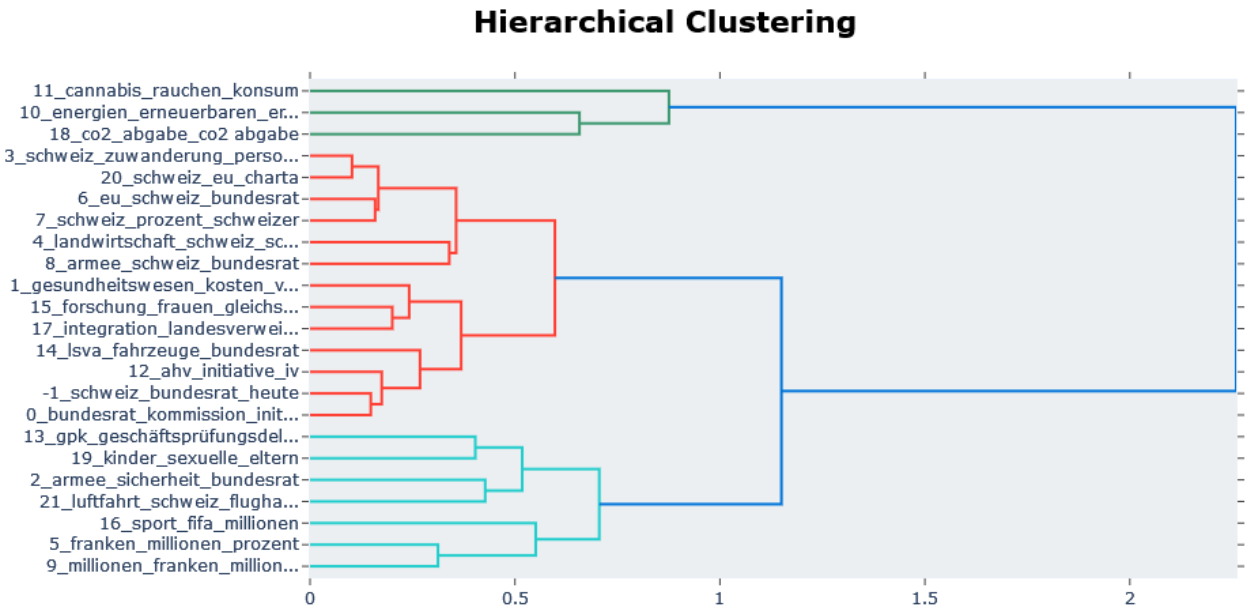
Visualize topics, their sizes, and their corresponding words

We saw in the previous slide that the topics and their corresponding words are more diverse and spread out, in the SP-speeches.

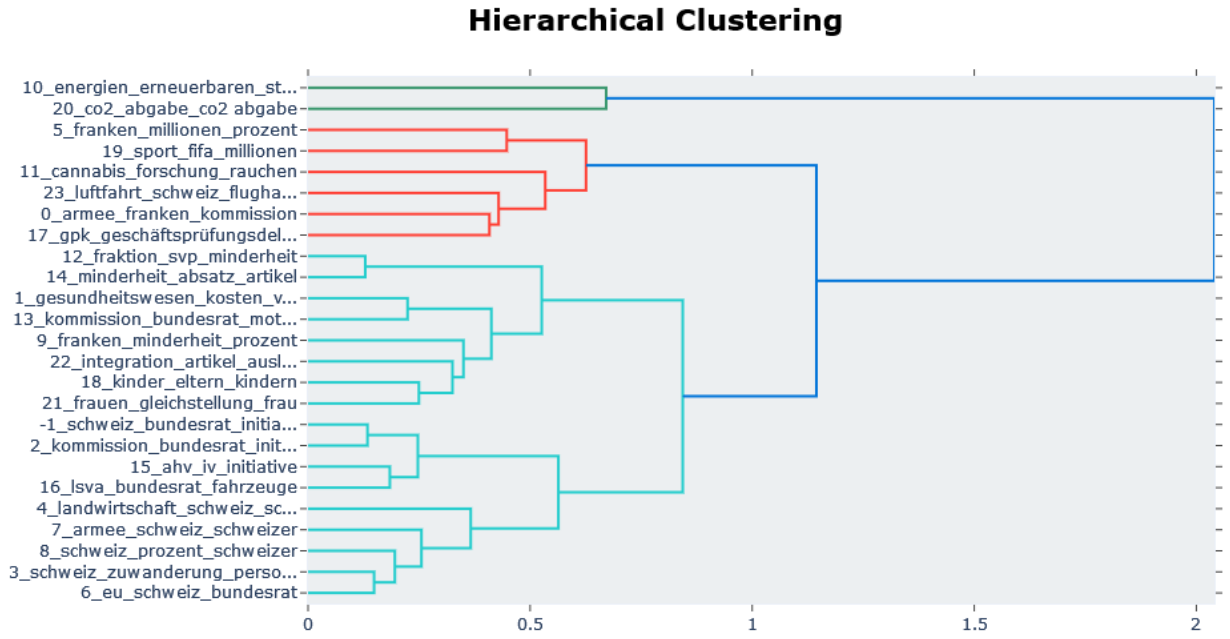
Whereas the SVP-speeches topics are more less spreaded.

Hierarchchical Clustering: Visualize Topic Hierarchy

SVP

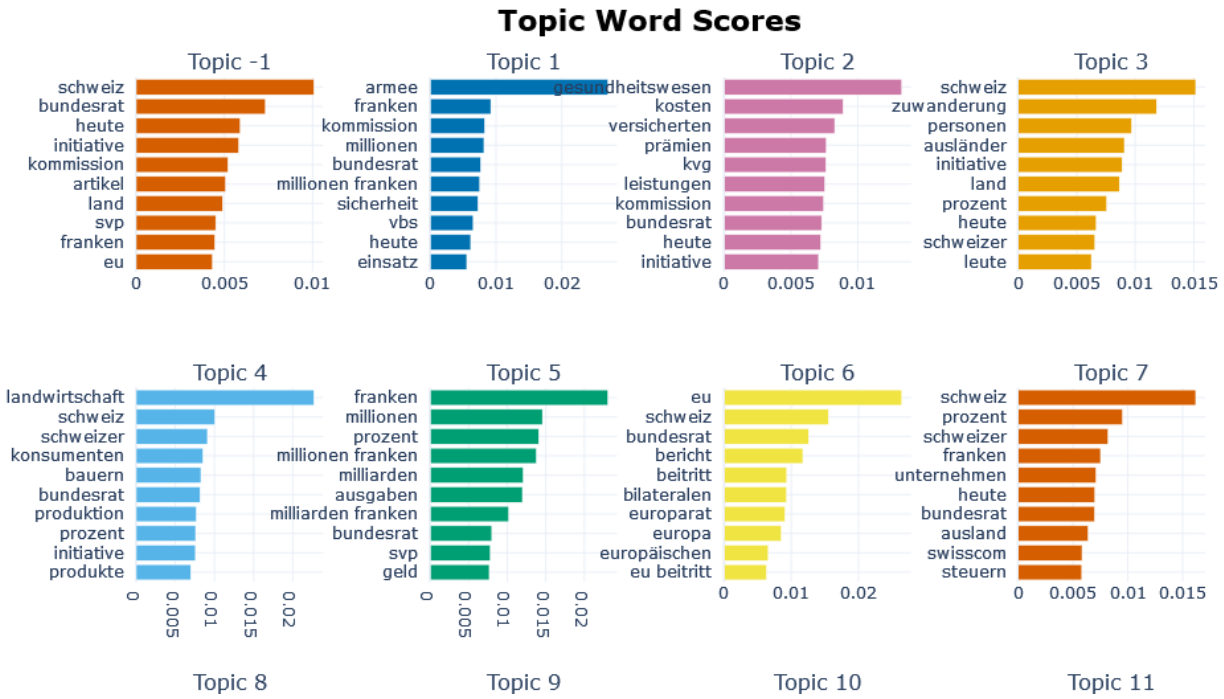


SP

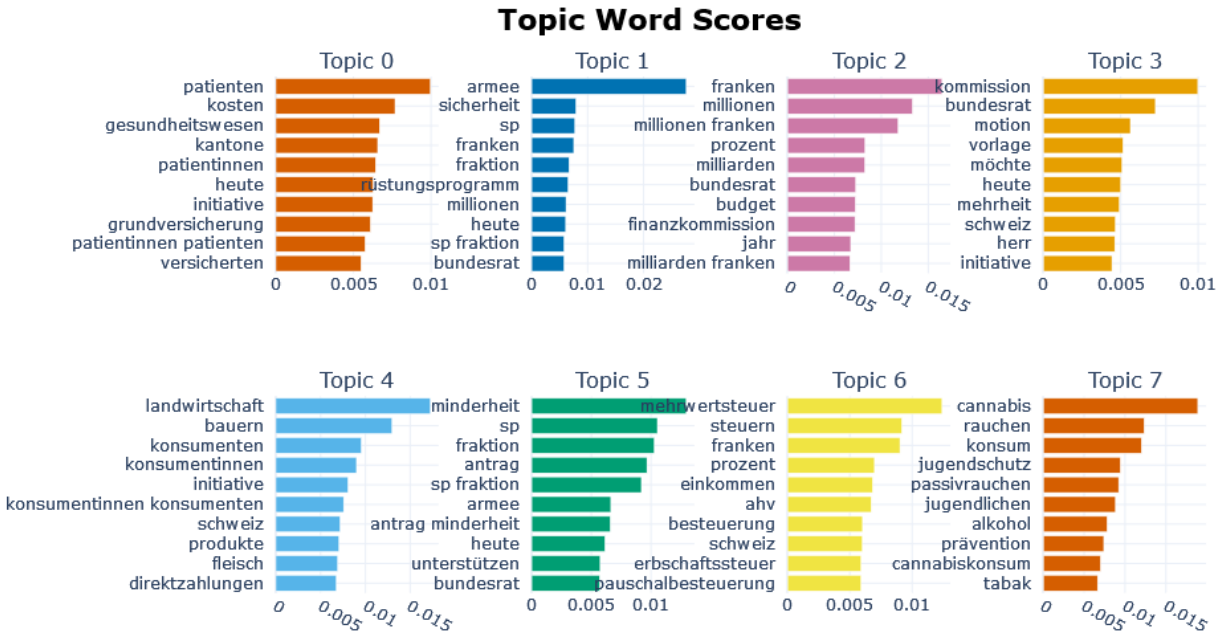


Visualize a barchart of selected topics

SVP

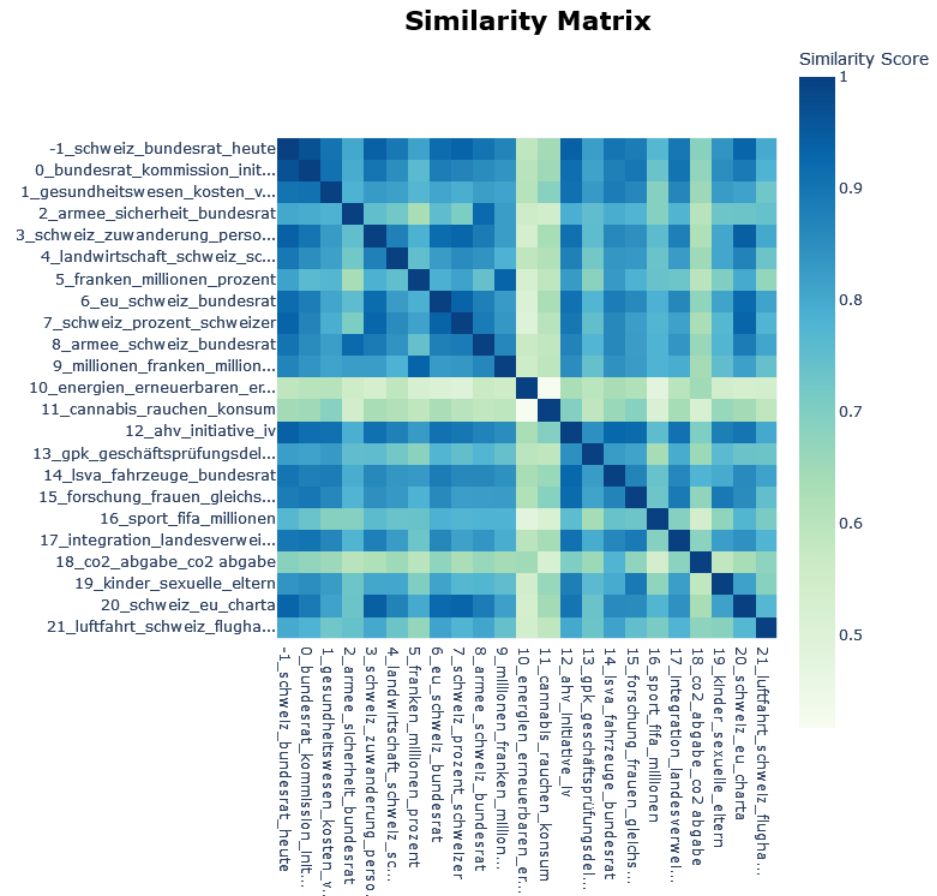


SP

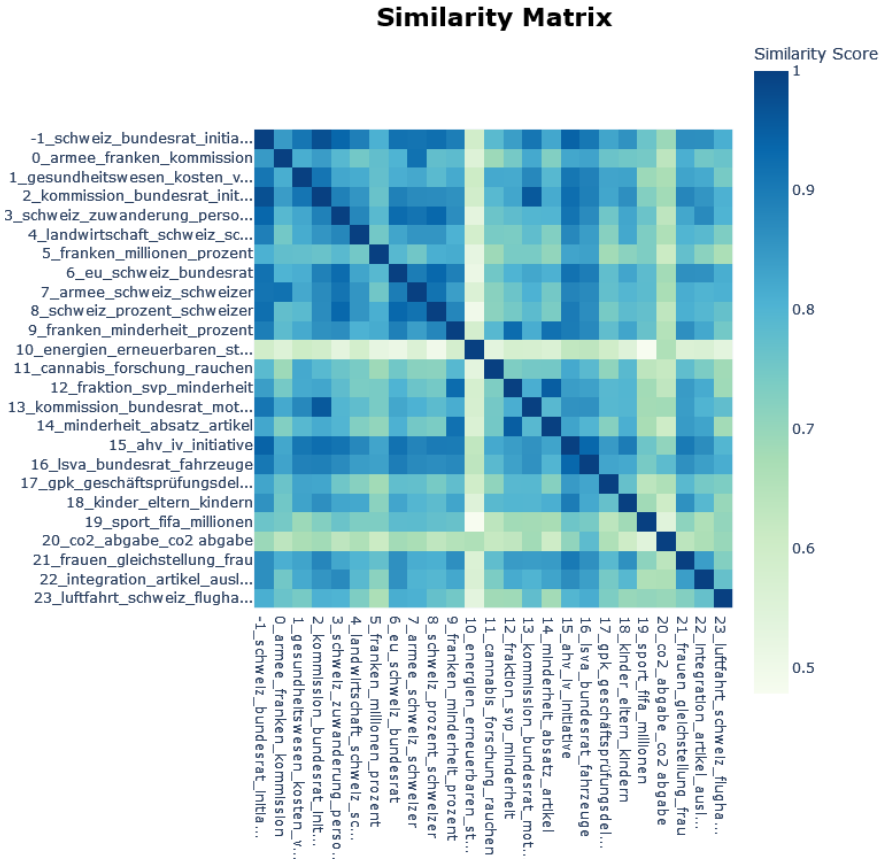


Visualize Topic Similarity

SVP



SP



How well performs the BERTopic? What keywords do you think have been used to filter the speeches?

- The keywords used in the German data are mainly nouns
- BERTopic performed pretty well but it does not take away the stopwords so one has to manually remove them or use NLTK.

Problems and difficulties

- Somewhat troubles interpreting the results. More than plotting and seeing the differences in the parties is not really possible. Maybe it would have been also more interesting to have the dates of the speeches in order to get the topics over time. This would make the data analysis somewhat more interesting and meaningful.