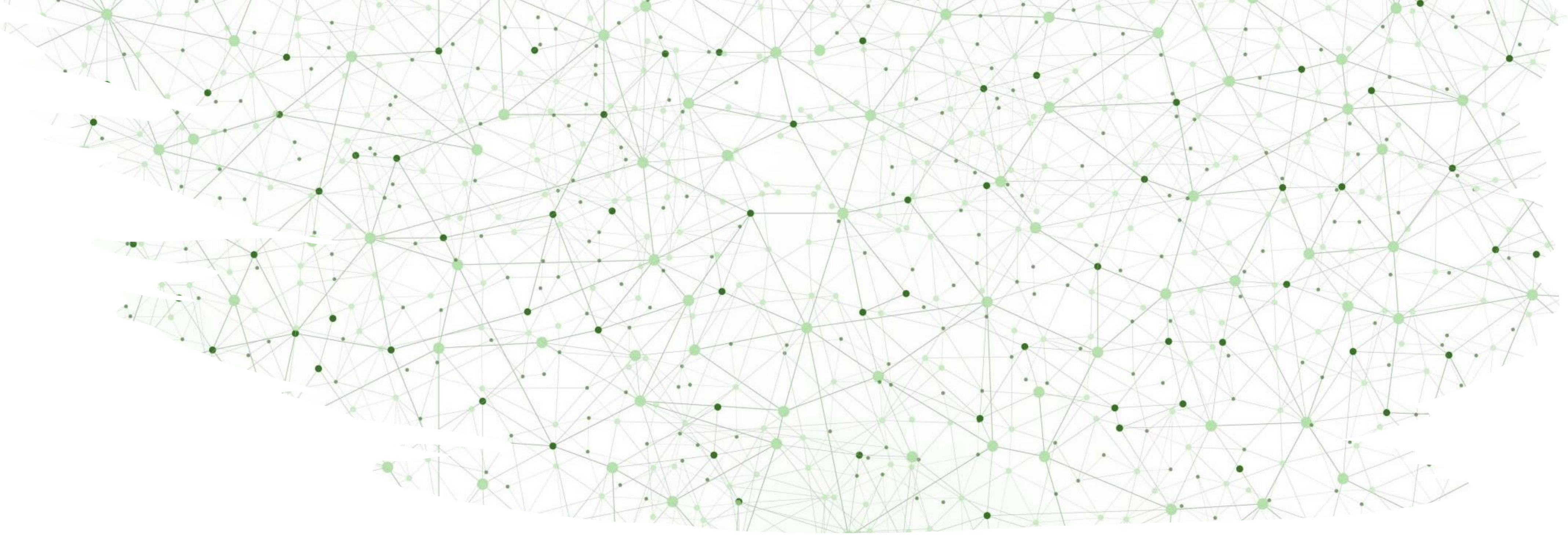
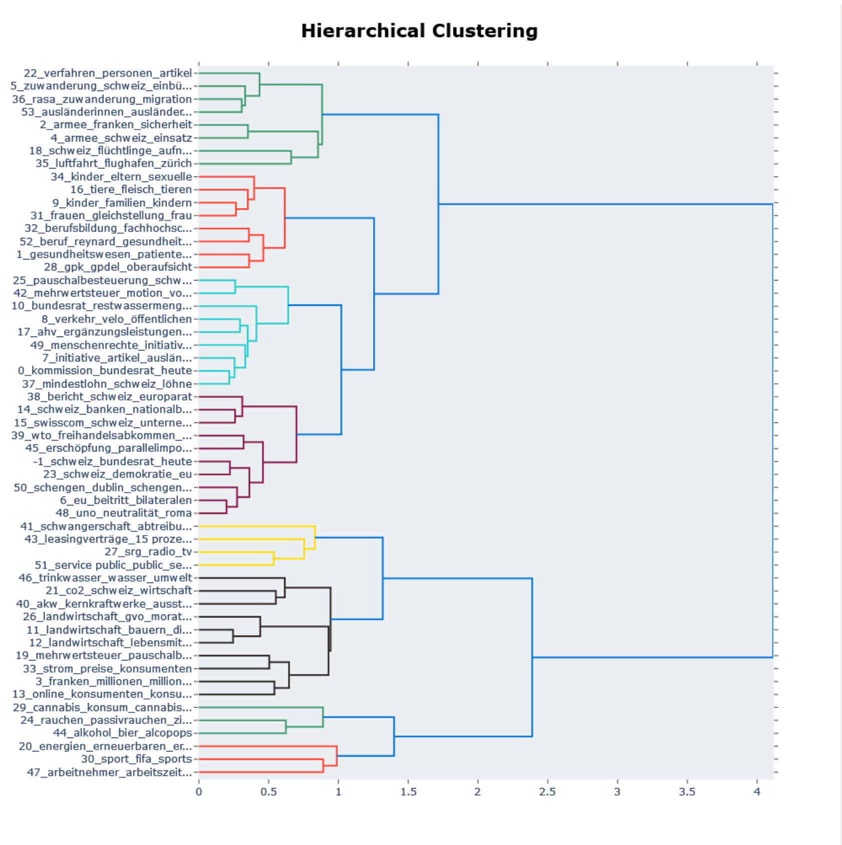
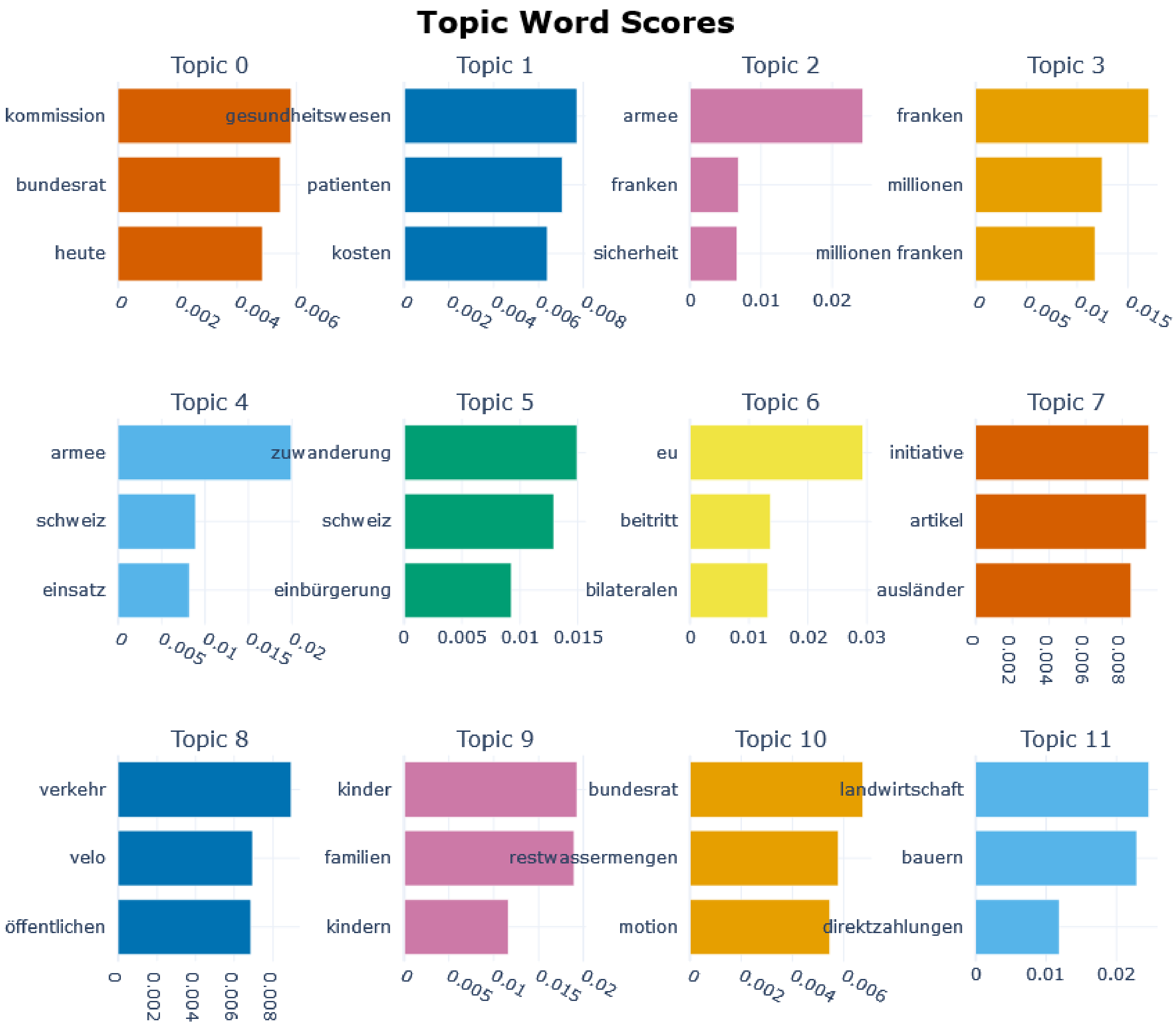


Report – Key Findings

TEXT MINING EXERCISE 3

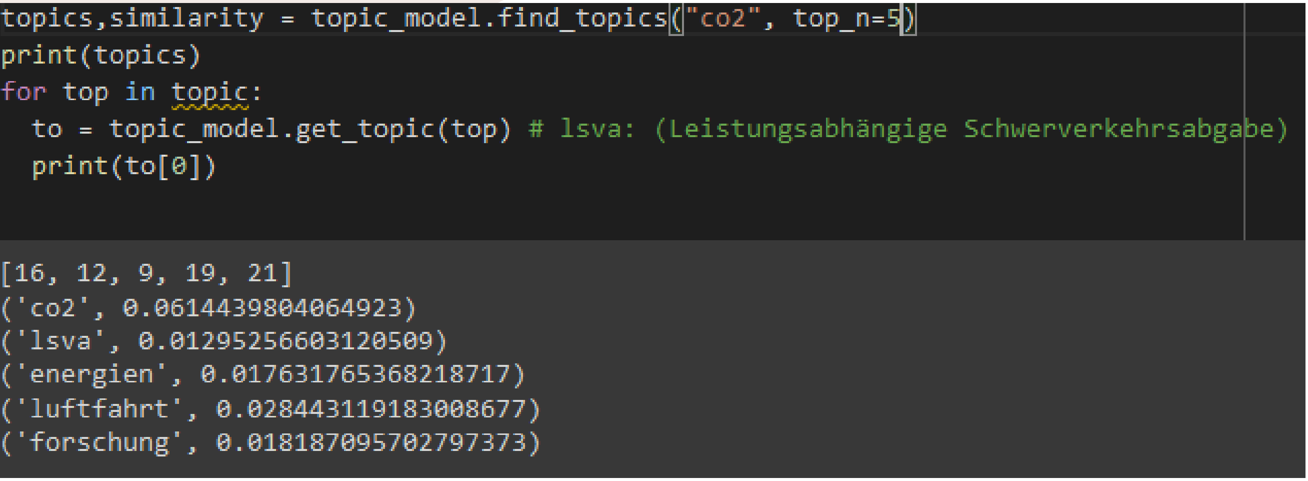
Full Data Analysis on the speeches


Hierarchical Clustering


Topic Word Scores

Looking for similar topics for ‘co2’

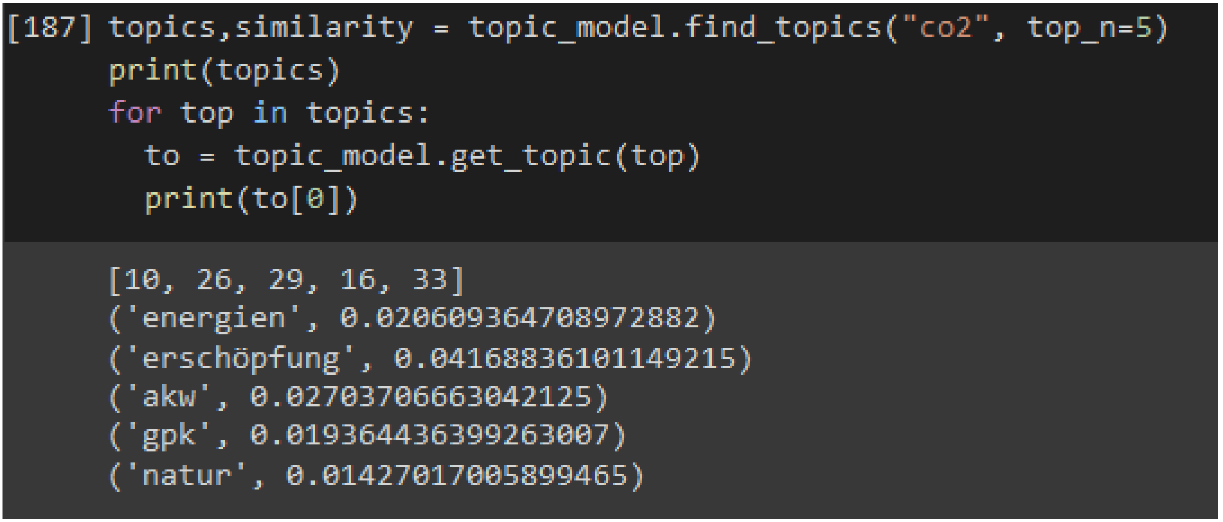
SVP



```
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



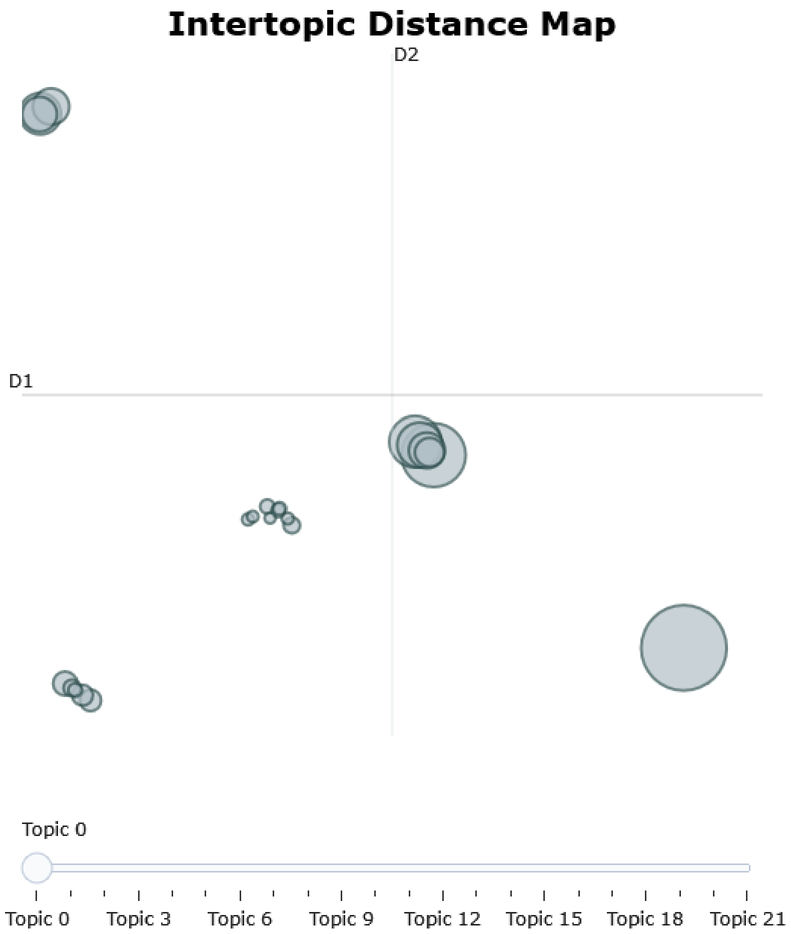
```
[187] topics,similarity = topic_model.find_topics("co2", top_n=5)
      print(topics)
      for top in topics:
        to = topic_model.get_topic(top)
        print(to[0])
```

```
      [10, 26, 29, 16, 33]
      ('energien', 0.020609364708972882)
      ('erschöpfung', 0.04168836101149215)
      ('akw', 0.02703706663042125)
      ('gpk', 0.019364436399263007)
      ('natur', 0.01427017005899465)
```

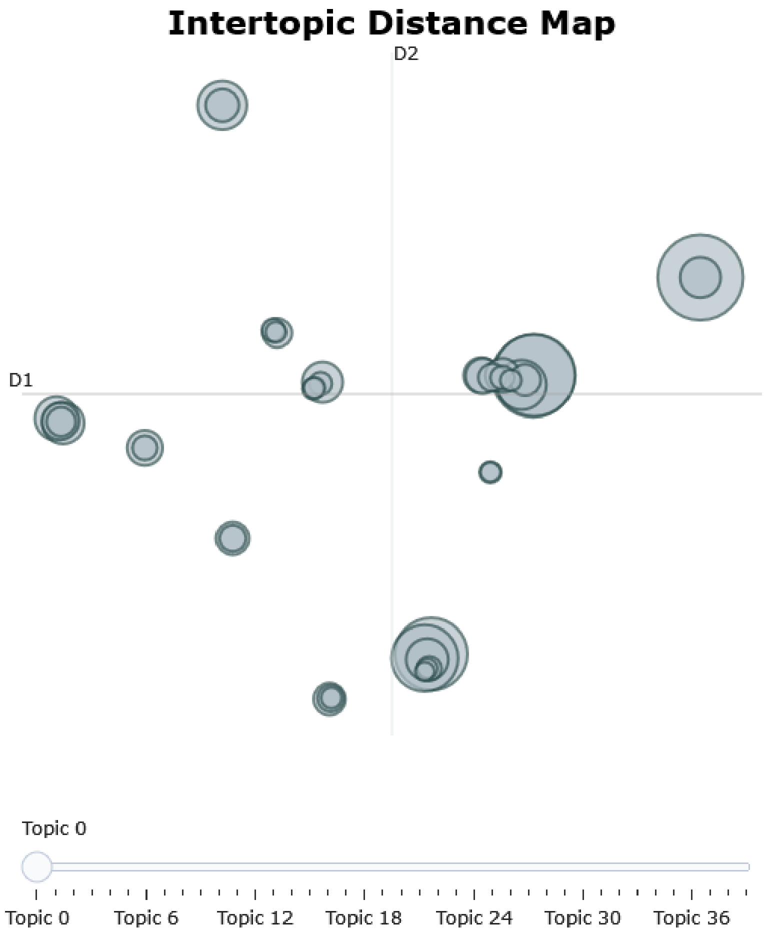
Interestingly, when we search for similar topics in the topic_model based on SVP-speeches vs. SP-speeches we clearly see a difference. It seems as if SVP ‘Co2’ related topics focus more on lsva(Scherverkehrsabgabe = shear traffic tax), luftfahrt (aviation) and forschung(research), which are topics more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



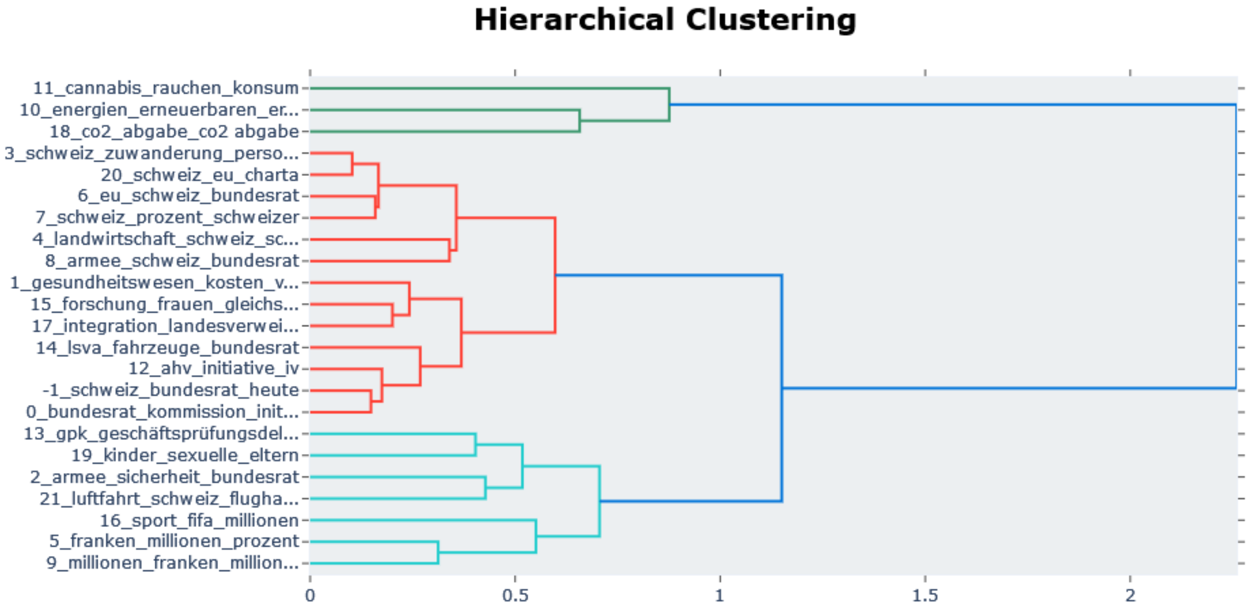
Visualize topics, their sizes, and their corresponding words

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

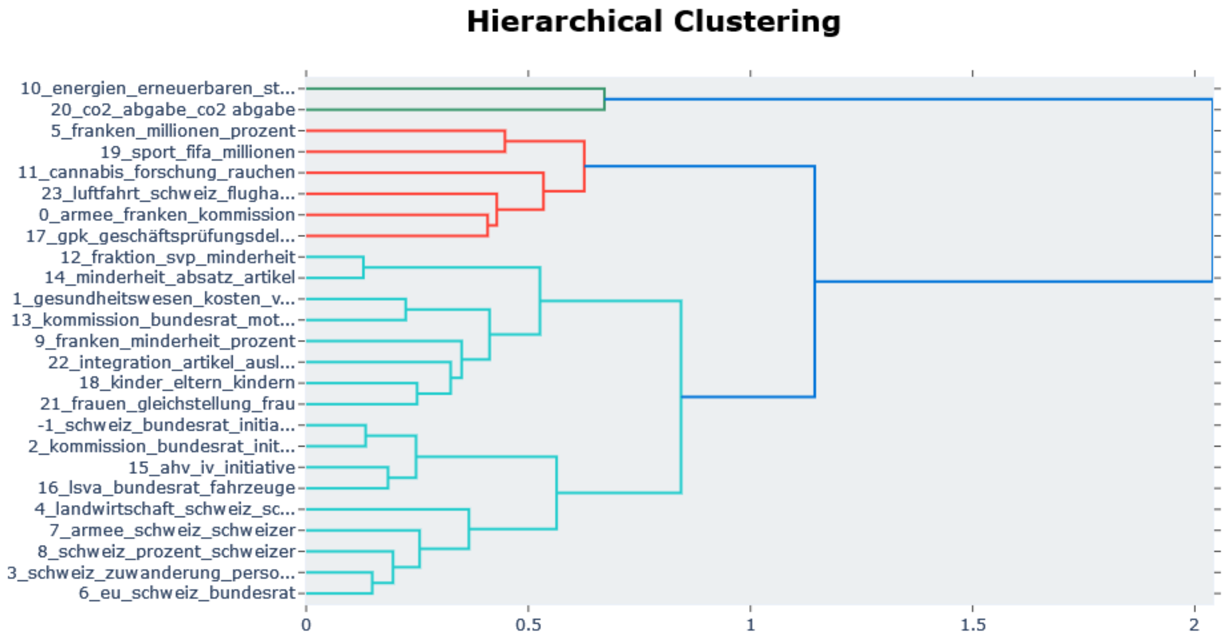
Whereas the SVP-speeches topics are more less spreaded.

Hierarchchical Clustering: Visualize Topic Hierarchy

SVP



SP



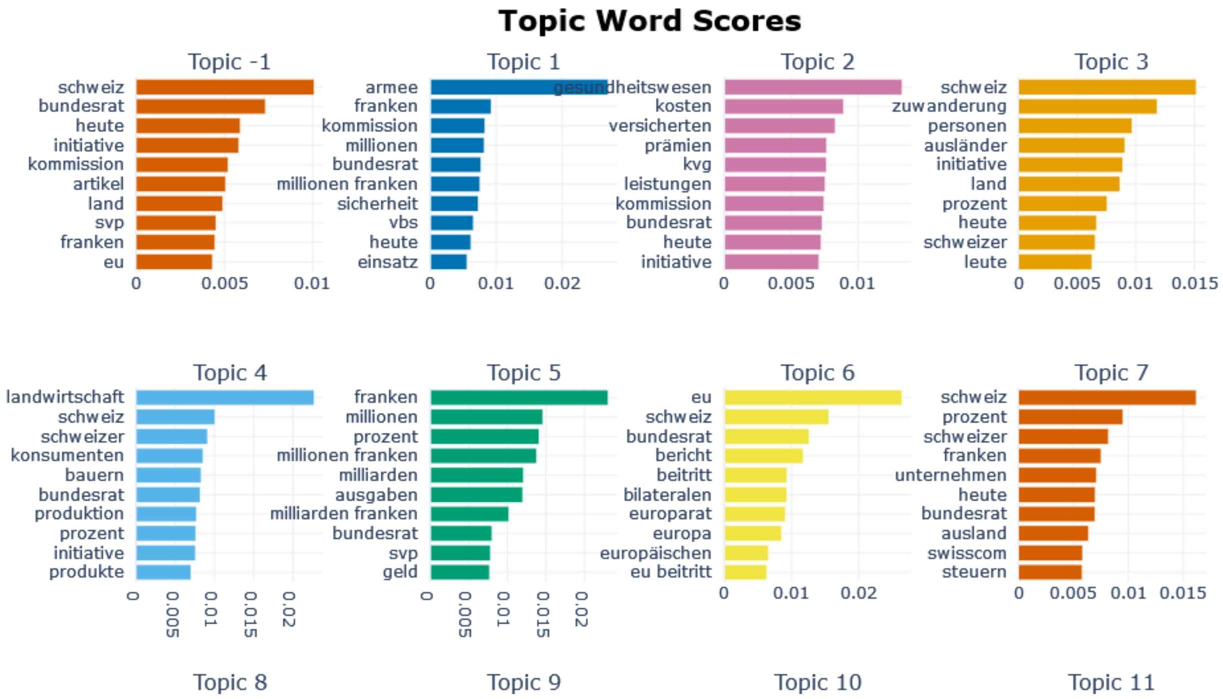
Hierarchical Clustering: Visualize Topic Hierarchy

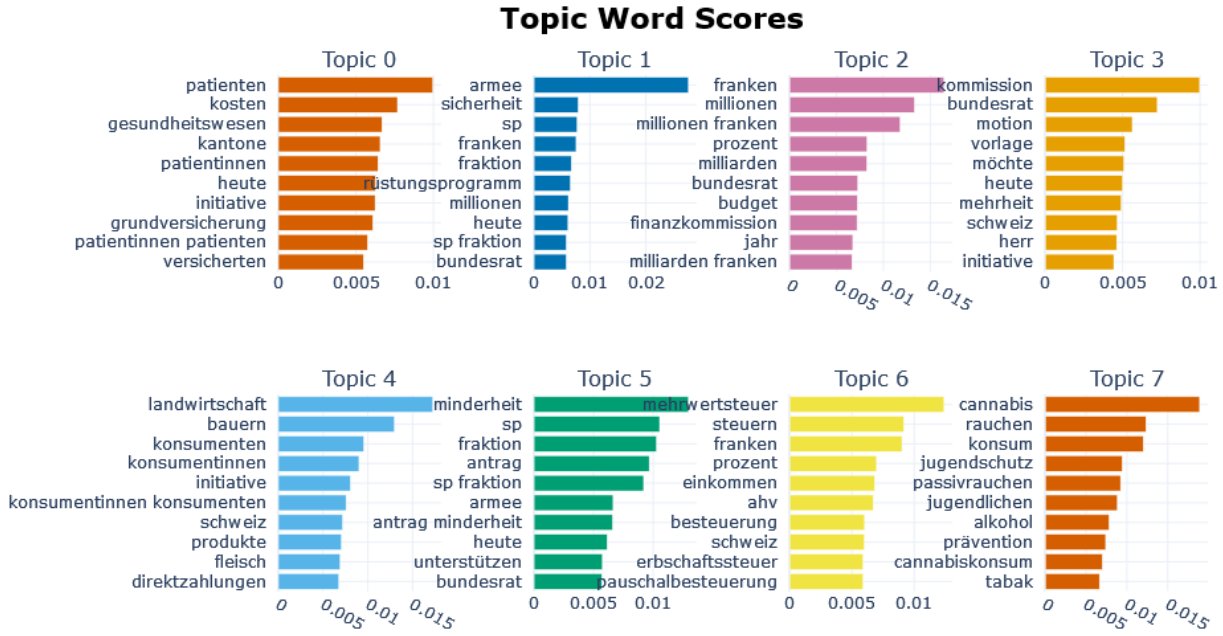
- We see a clear difference in the clustering of the topics.
- But it is clear, depending on the cut of the clustering tree, we can see a grouping of 3 colors (red, green, light blue).
- The green group seems to be about energy/co2 in the SP-speeches but according to the analysis on SVP-speeches also the cannabis consumption topic is a part of that group. This makes the interpretation more difficult.
- Also the interpretation of the red grouping is difficult, as it seems to be about money-related topics in SP-speeches it is very different in the SVP-speeches, maybe about initiatives (Gesundheitswesen/Gleichstellung/AHV/cannabis).
- Then the light blue topic group in the SVP-speeches it seems to be more about money-related topics and in the SP-speeches more about initiatives.
- If we look into the groups and pick for example, the topic 5 it falls in both cases SVP/SP-speeches under a group related to money.
- According to the title and the group if falls into, topic 11 in SVP-speech in my opinion is quite a misfit.

Visualize a barchart of selected topics

SVP

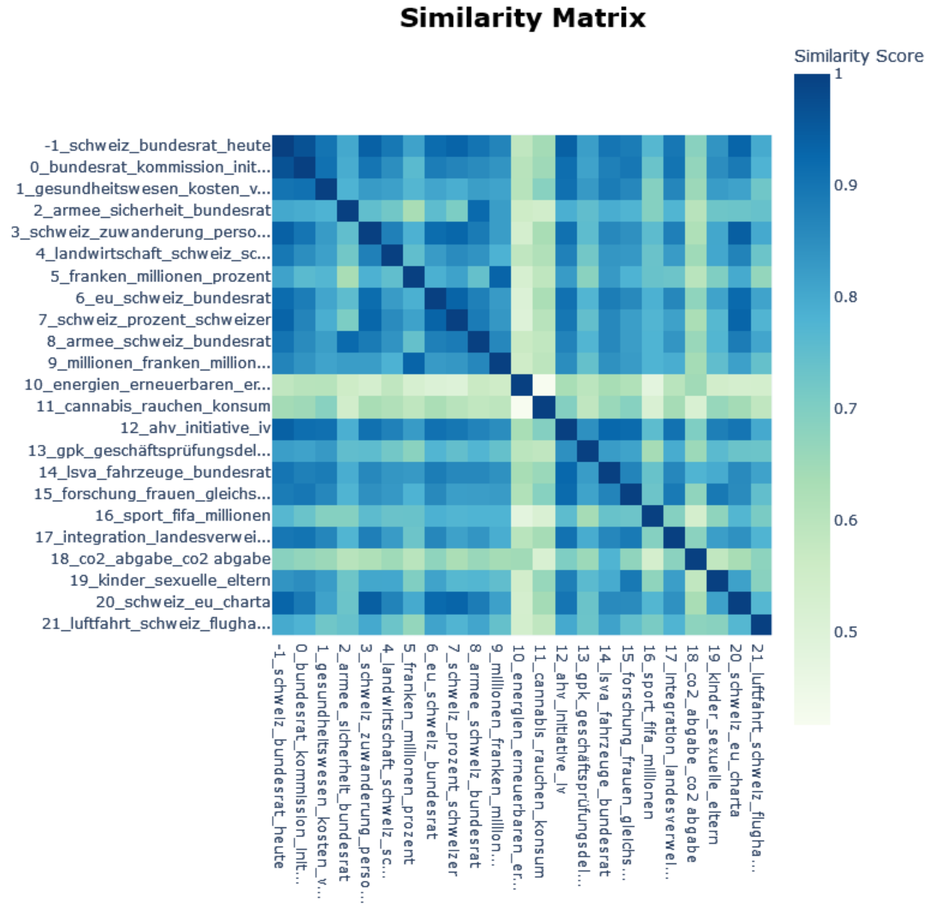
SP



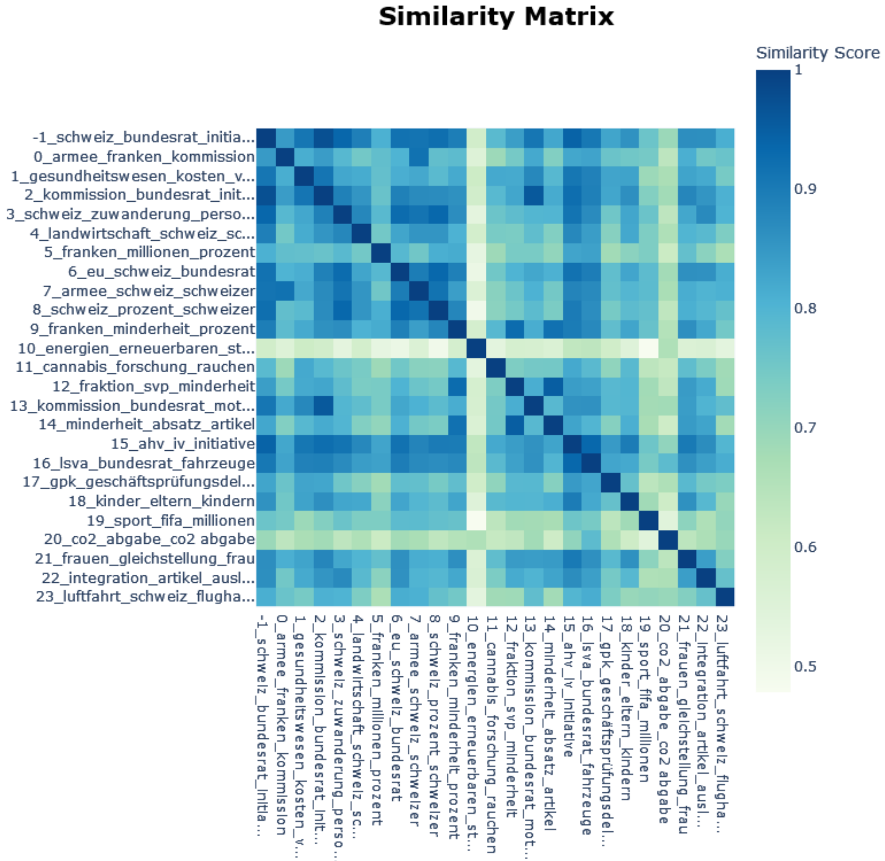


Visualize Topic Similarity

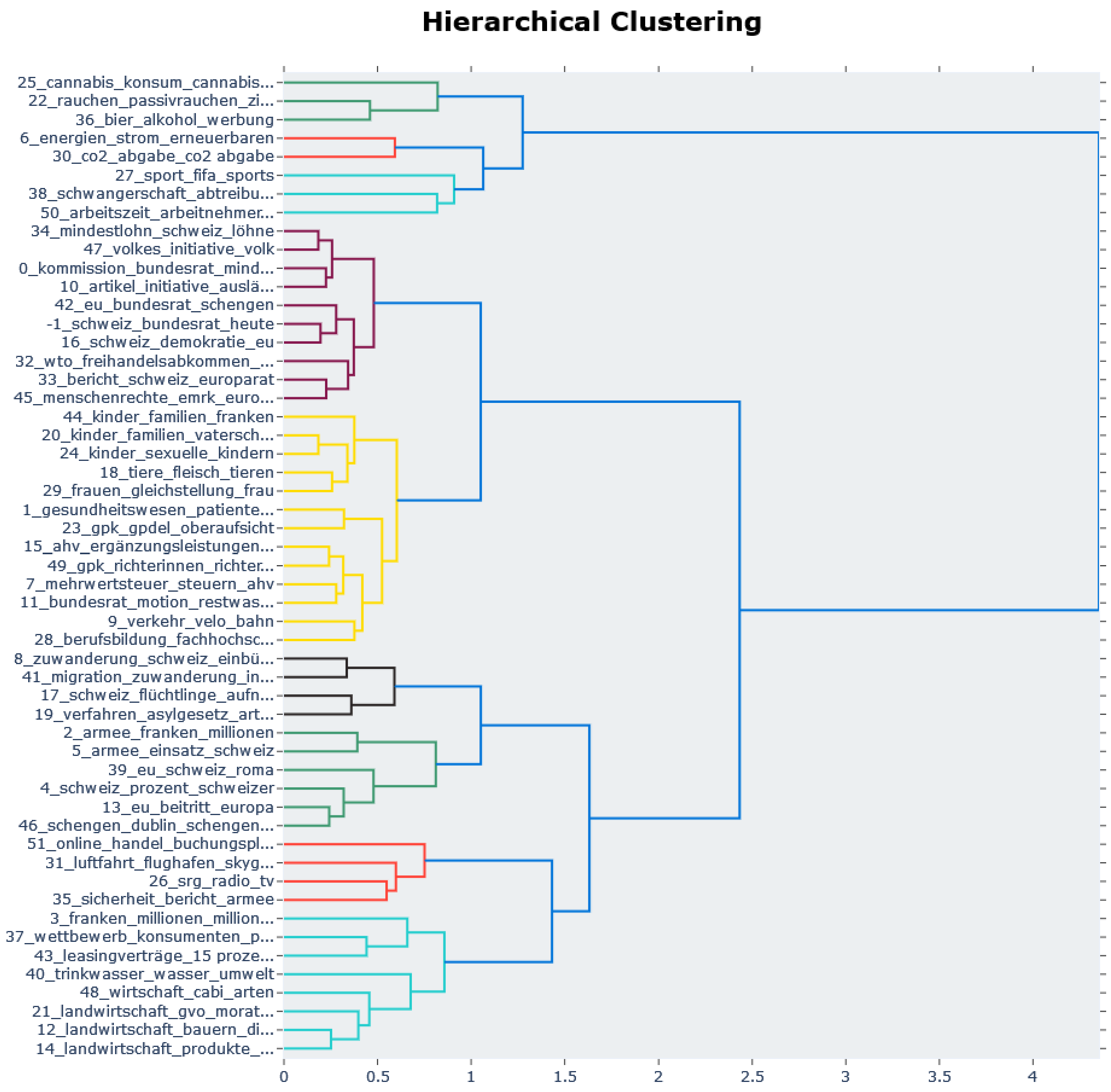
SVP



SP



- In the full dataset we see, depending on the cut of the tree, 7 groups. My interpretation of the groups would be:
    - Migration, energy, drugs, pregnancy vs. work, initiatives, money and economy, agriculture,..
    - Again, the interpretation of the grouping seems to be difficult. Why is ‘verkehr_velo_bahn’ topic 9, together grouped with ‘kinder familien und vaterschaft’ topic 20 ?
    - And why in the green group is passive smoking and cannabis not first grouped but passive smoking is first clustered with beer and alcohol advertisement.



Hierarchical Clustering

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

- 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.
- It took about a day.

Problems and difficulties

- Somewhat troubles interpreting the results, especially the ones from the hierarchical clustering . 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.
- As we saw that the interpretation of the clustering was in some cases rather difficult. Especially if the groups are too large.
- I would suggest that there are not too many topics for the German data, as the grouping/clustering gets more difficult and less easy to interpret. Moreover for each topic in the topic-analysis I would not look at more than 10 words, best would be 3 just in order to get the main idea of a topic.