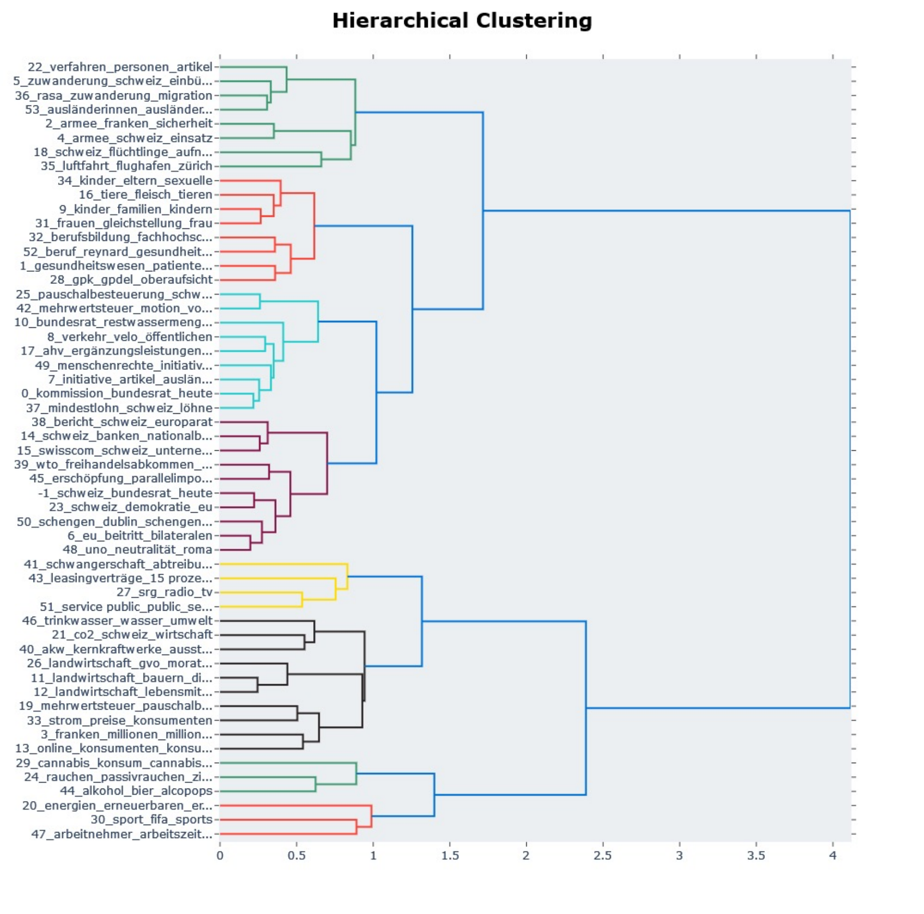
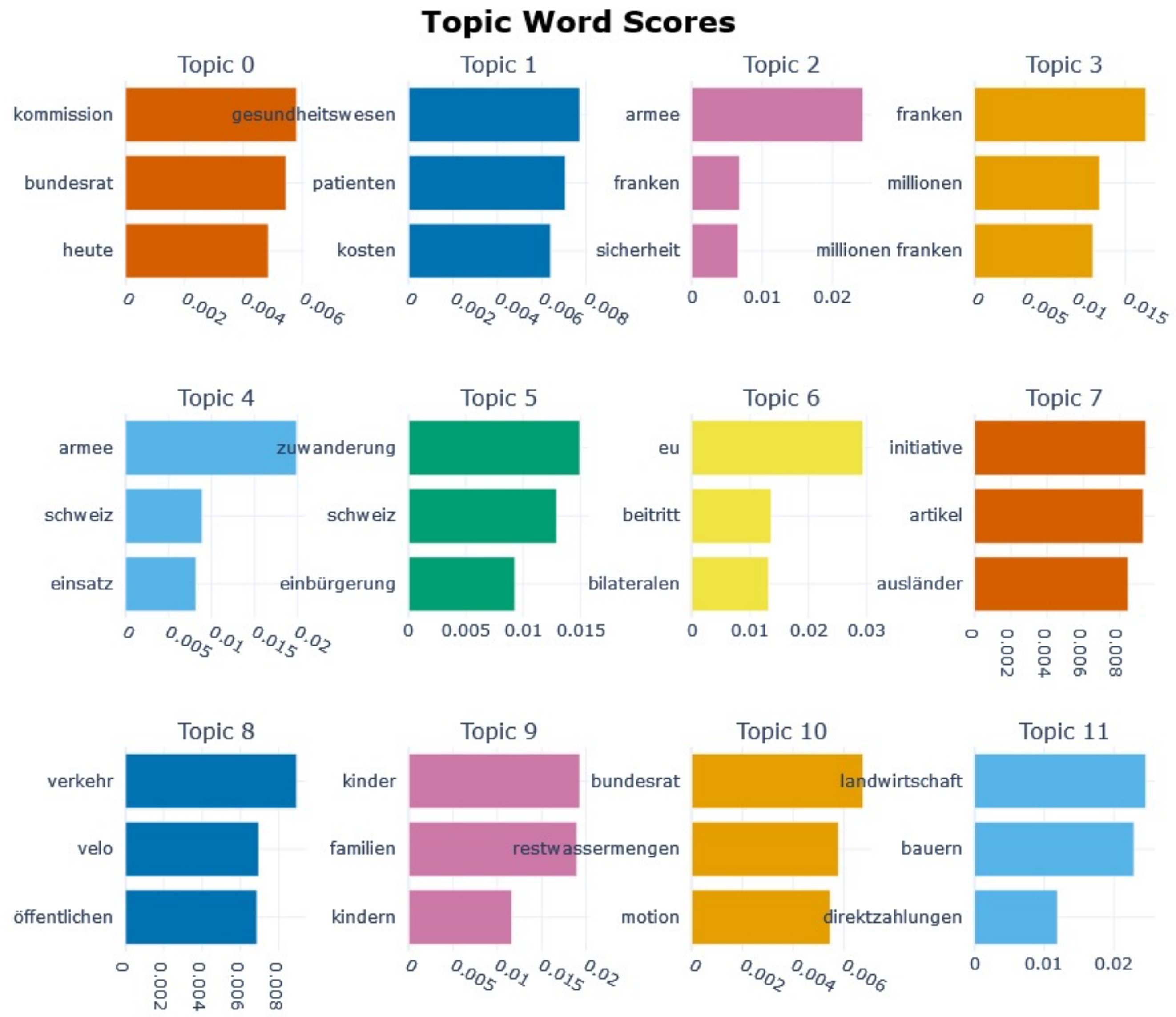
Report – Key Findings

TEXT MINING EXERCISE 3

REBECKA FAHRNI & JESSICA ROADY

Analysis of the speeches





Looking for similar topics to 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

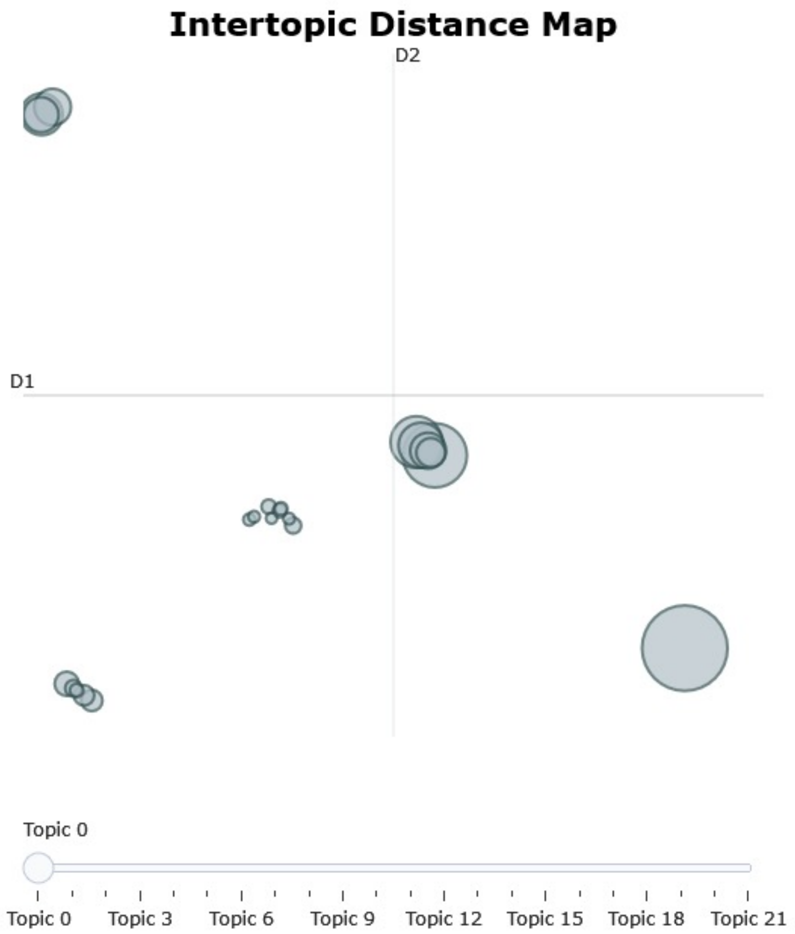
```
[187] topics,similarity = topic_model.find_topics("co2", top_n=5)
      print(topics)
      for top in topics:
        to = topic_model.get_topic(top)
        print(to[0])
```

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

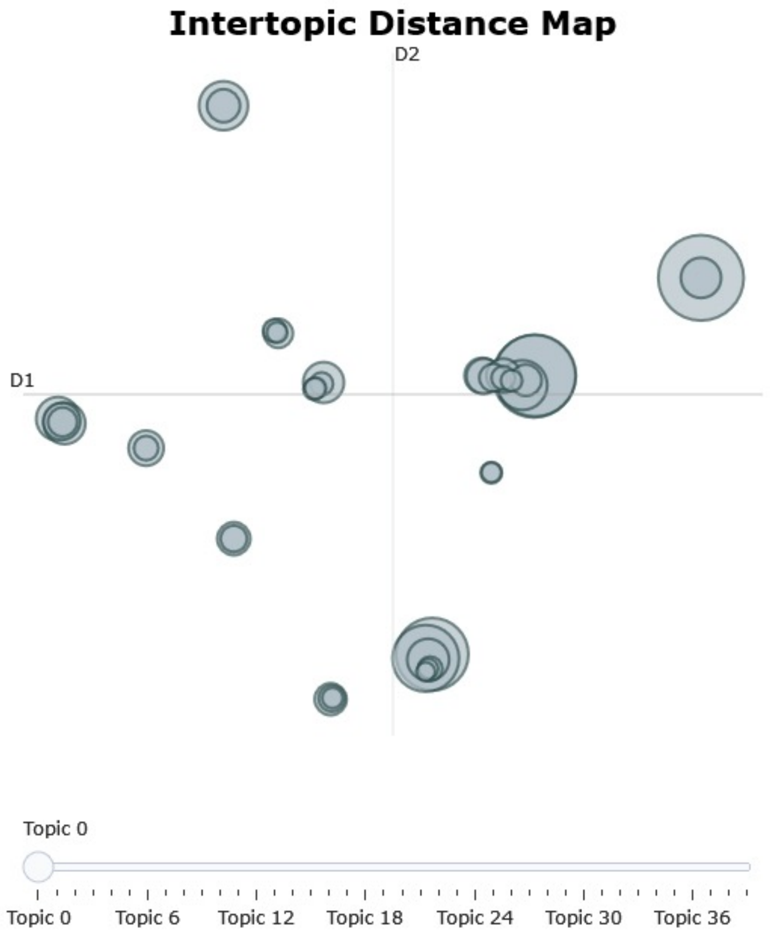
Interestingly, when we search for similar topics in **topic_model** based on SVP vs. SP speeches, we clearly see a difference. It seems as if SVP **co2**-related topics focus more on **lsva** (*Scherverkehrsabgabe* = sheer traffic tax), **luftfahrt** (aviation) and **forschung** (research), which are more economy-related topics. Topics similar to **co2** in the SP speeches are very different – for example, **erschöpfung** (exhaustion) here is closely related to Co2.

Visualize topics, sizes, and corresponding words

SVP



SP



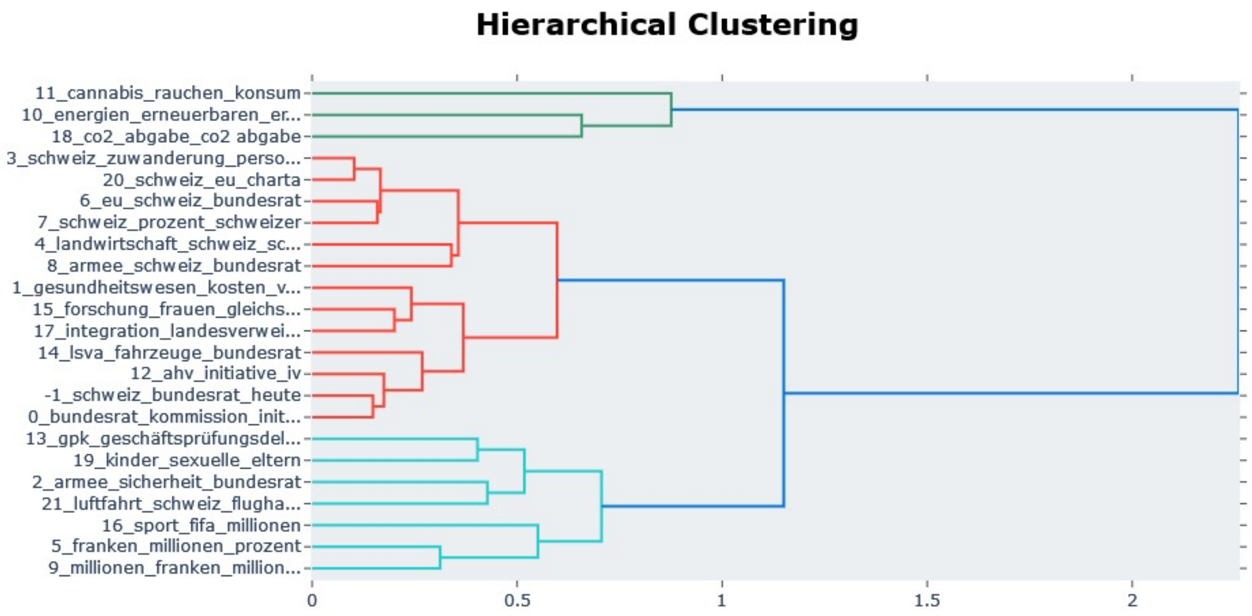
Visualize topics, sizes, and corresponding words

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

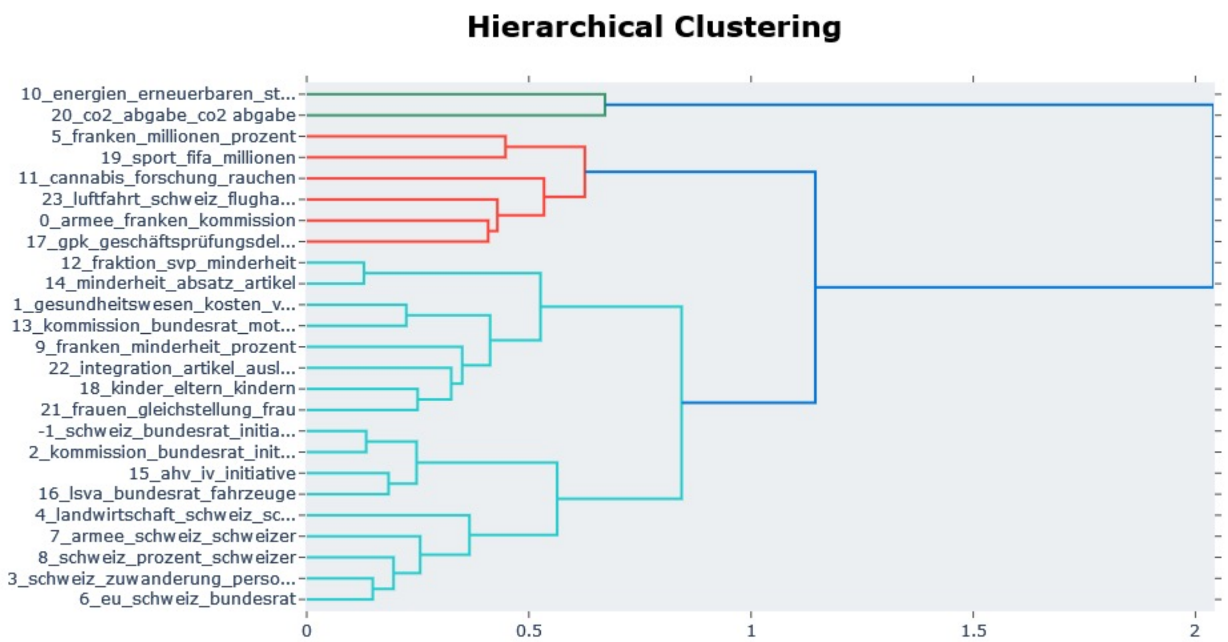
Topics in the SVP speeches are more concentrated.

Hierarchical Clustering: Visualize Topic Hierarchy

SVP



SP

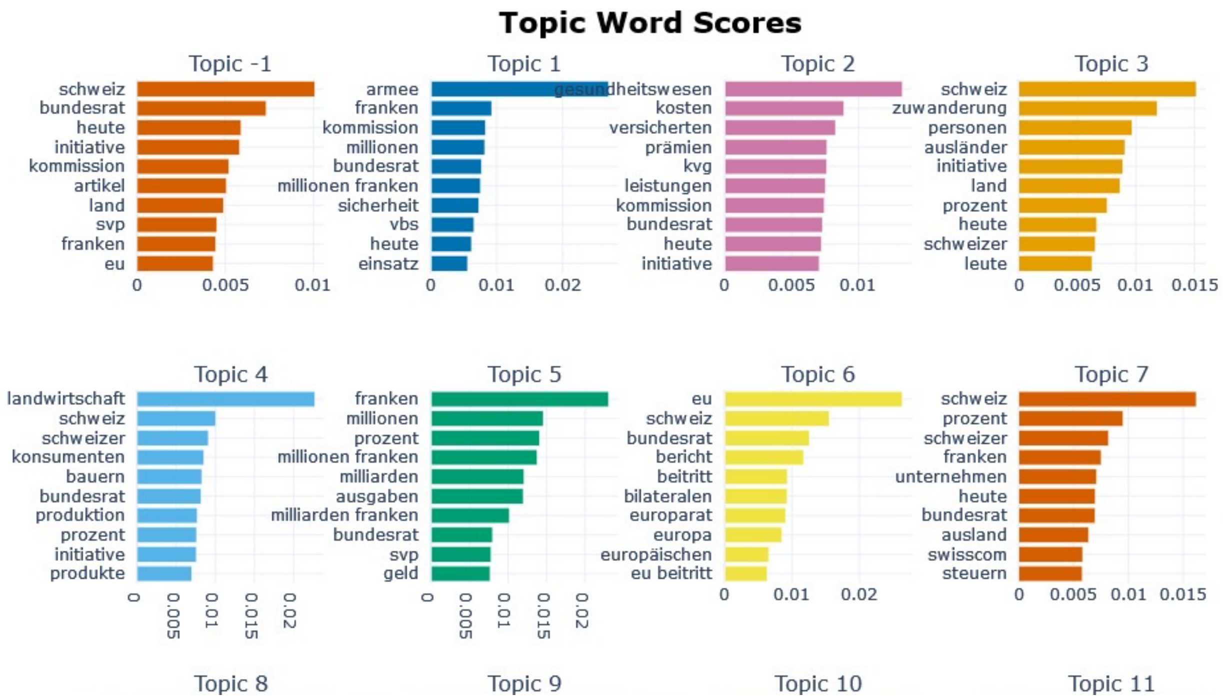


Hierarchical Clustering: Visualize Topic Hierarchy

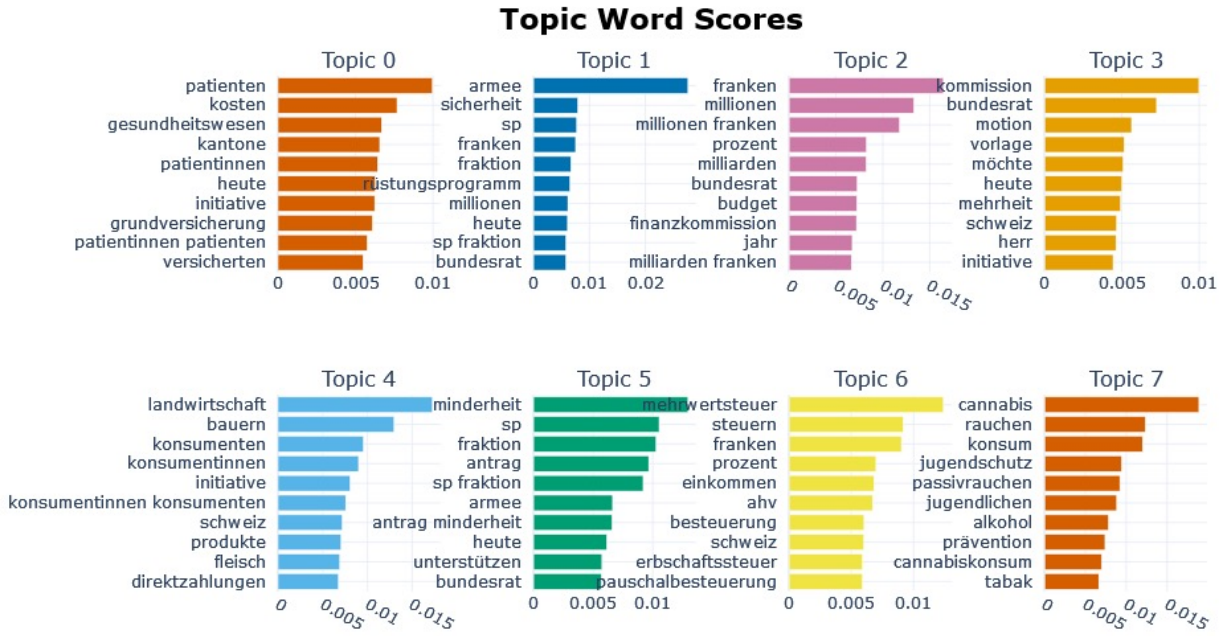
- We see a clear difference in the clustering of the topics.
- By looking at the cut of the clustering tree, we can see a grouping of 3 colors (red, green, light blue).
- The green group seems to be about energy/CO2 in the SP speeches, but according to the analysis of SVP speeches, the topic of cannabis consumption is also a part of that group. This makes the interpretation more difficult.
- Interpretation of the red group is also difficult, as it seems to be about money in SP speeches but something else entirely in the SVP speeches (perhaps initiatives – Gesundheitswesen/Gleichstellung/AHV/cannabis).
- Conversely, the light blue topic group in the SVP speeches seems to be more about money and in the SP speeches more about initiatives.
- If we examine the groups and pick, for example, topic 5, it falls under a group related to money in both SVP and SP speeches.
- According to the title and the group it falls into, topic 11 in SVP speeches seems a misfit.

Visualize a barchart of selected topics

SVP

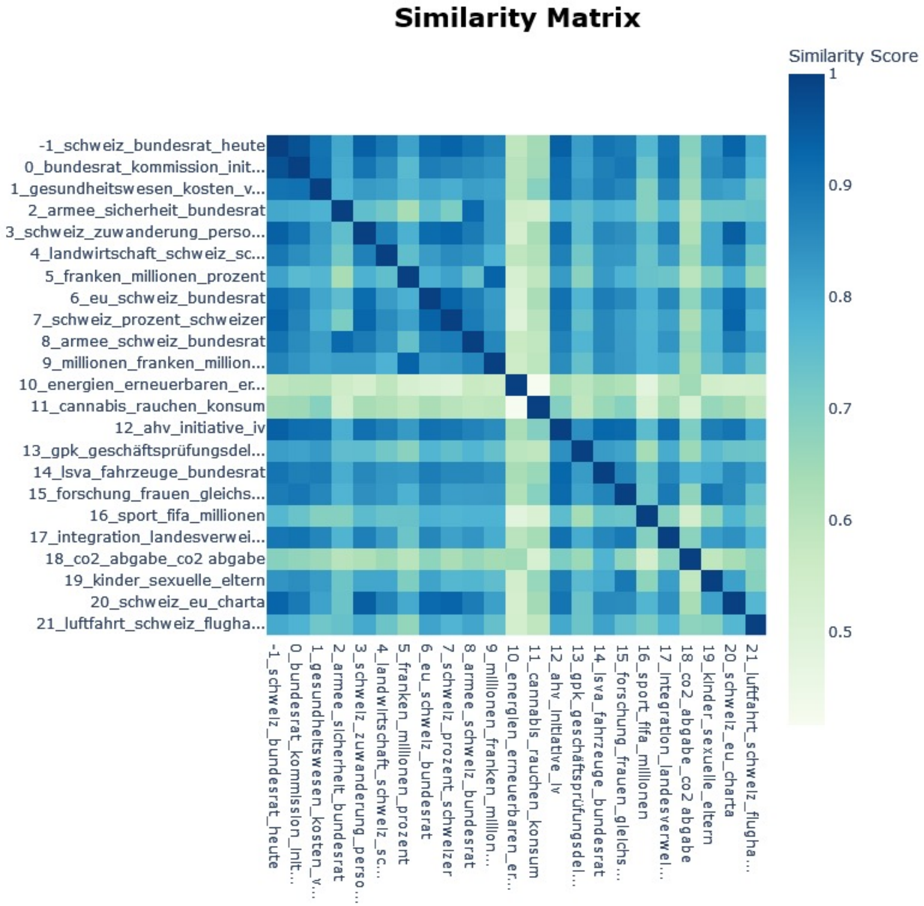


SP

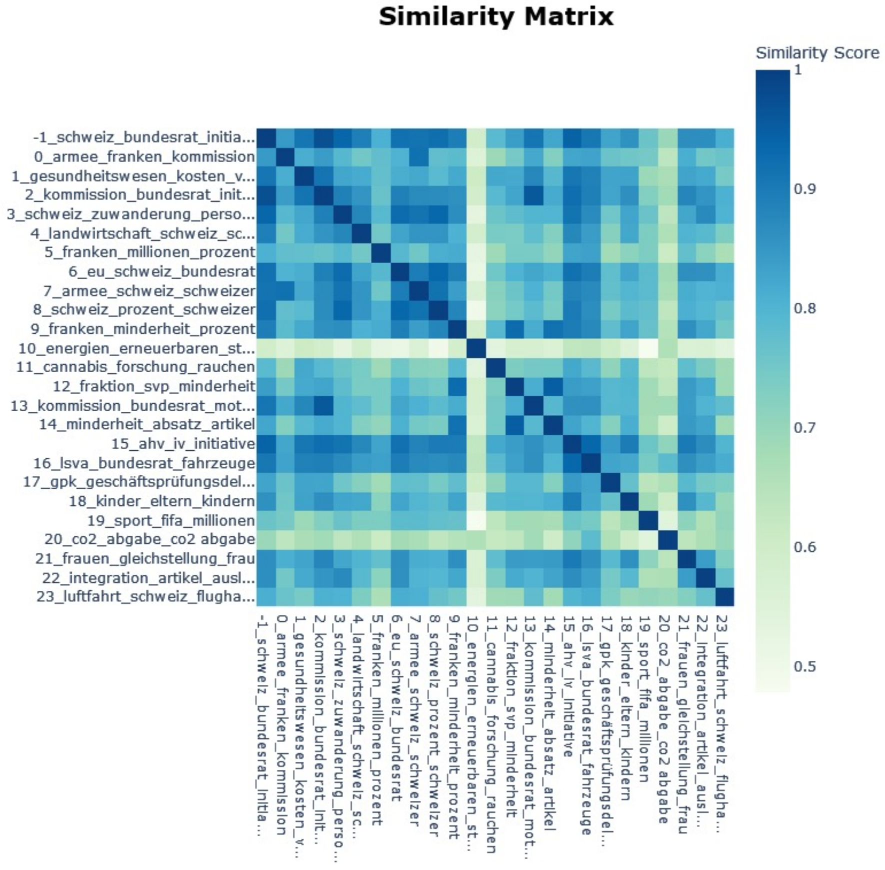


Visualize Topic Similarity

SVP



SP

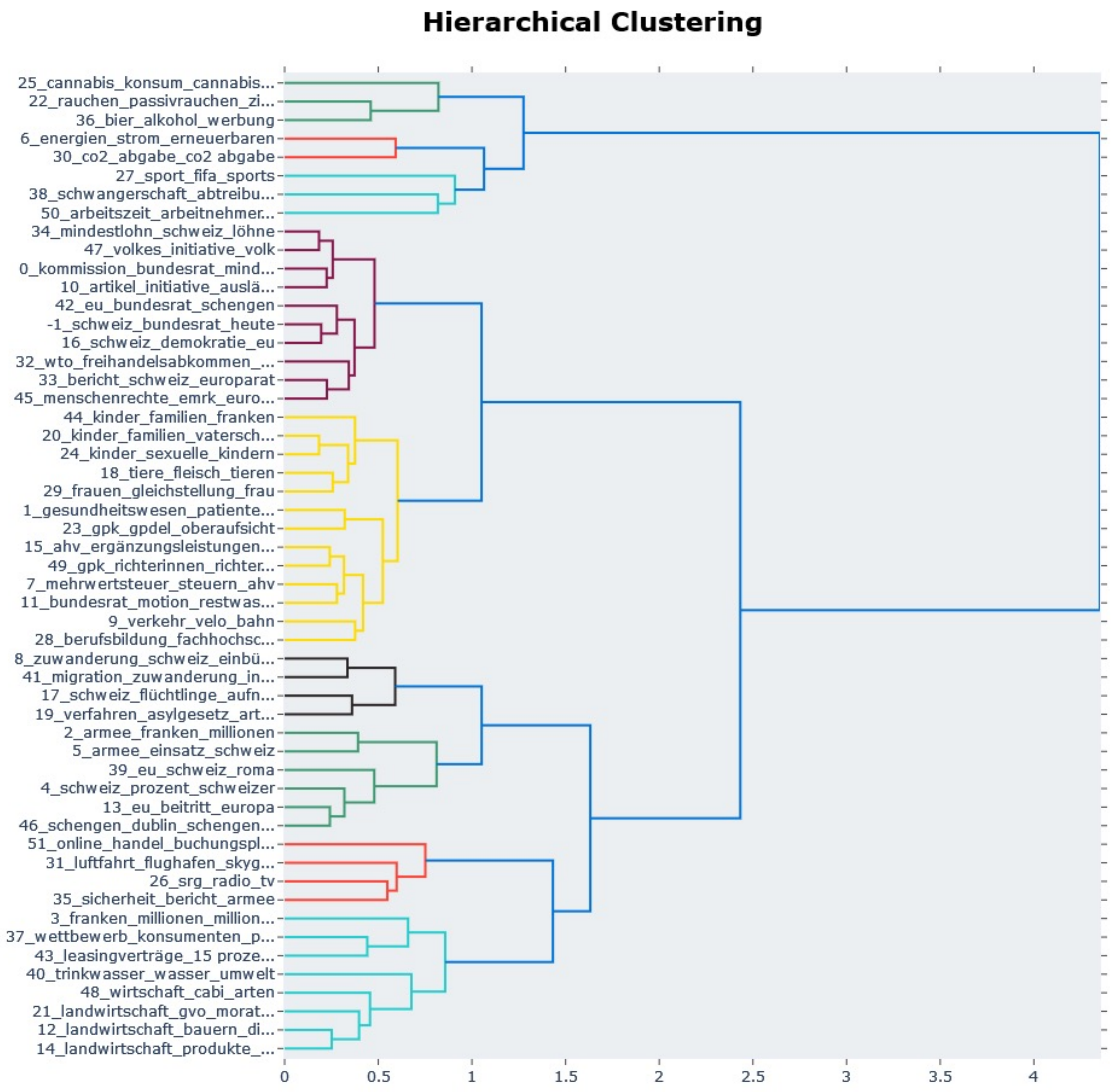


In the full dataset we see 7 groups, depending on the cut of the tree. Our interpretations of the groups would be:

- 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) grouped with **kinder_familien_und_vaterschaft** (topic 20)?
- In the green group, why is passive smoking and cannabis not first grouped, but passive smoking is first clustered with beer and alcohol advertisement?



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

- The keywords used in the German data are mainly nouns.
- BERTopic performed pretty well, but even with 5000+ data points does not ignore stopwords.
- It took about two days.

Problems and Difficulties

Code

- large amount of data needed before stopwords disappear
- topic -1 is annoying

Interpretation

- trouble interpreting results, particularly from hierarchical clustering
- more than plotting and seeing the differences between parties is not really possible
- preserving the dates of texts would allow for diachronous modelling of topics
- limiting the number of topics and keywords per topic improves intelligibility