Battle Plan:

1. Approach similar to “Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora”

* No choice of topic necessary
  + Use PCA to get to 2-dimensional space
* Greater computational/training effort
* Potentially hard to find good prompts (as we are not focussing on a specific topic)

1. Manifesto-based approach

* Choose a topic with Positive/Negative Dimension
  + 1-dimensional space
* Use their own (context) model to get scores per sentence
* Aggregate scores of each sentence for whole document
* Probably easy to find prompts as we have clear topics
* Identify Topic
  + Based on data availability and Model availability
  + Candidates:
    - Covid (Less/More Restrictive)
      * No Manifesto Dimension
    - Ukraine (Pro/Contra Military Aid)
      * No Manifesto Dimension
    - Immigration and Integration
      * <https://manifesto-project.wzb.eu/information/documents/pimpo>
      * <https://manifesto-project.wzb.eu/coding_schemes/mp_v4>
        + Basically all Variables with Positive/Negative can be used for spatial representation
* Choose Spatial Model(s)
  + Candidates:
    - Sentence-BERT Model + Dimesionality Reduction + Clustering
      * Buryakov, Hino, Kovacs, Serdült (Uni Zürich)
    - MisterBertGoesToParliament
      * Couldn’t find model/paper
      * <https://coms.events/epsa-2022/data/abstracts/en/abstract_0299.html> (starting point to search further)
    - Manifesto-Project (Context Model): <https://huggingface.co/manifesto-project/manifestoberta-xlm-roberta-56policy-topics-context-2023-1-1>
      * There is also a Sentence Model
        + In both cases, we probably have to map the texts sentence by sentence and aggregate it! (But I think with context is potentially preferable, while a bit harder to code)
      * This gives us probabilities for all(!) the categories in the Manifesto Project (sorted)
        + We could focus on topics where we have positive and negative (e.g., Internationalism, European Community/Union)
      * Model can handle German texts
      * No need to focus on specific contexts!
    - <http://www.wordfish.org/>
      * Super simple statistical model of word counts
      * Good baseline
    - WordScores
      * Focuses on word frequency across documents (with known ratings), but there are two scores (document scores and wordscores)
        + Assumption: all words have the same informative value
      * Reference texts and unseen texts values can’t be compared directly (rescaling necessary as the variance is understated)
      * Empirically well-proven performance
* Data sets
  + ParlEE plenary speeches data set: Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZY3RV7>