You should limit the length of your answers as indicated in the questions!
Not following these limits will result in deduction of points!

Questions (1), (2), and (3) are related to each other and you may iterate over those three questions together to improve your results. In all three questions you will be working with the dataset "debates_2022.csv" (available in TeachCenter), which includes transcripts of all talks in the European parliament in 2022 with some additional metadata. All talk transcripts are in English. Your goal in questions (1), (2), and (3) is to extract the most important topics of these talks by clustering the talks.

1. *Feature Engineering.* Extract the features from the talk transcripts by computing tf-idf scores for words. You can use TfIdfVectorizer. Read the documentation of the vectorizer carefully and decide on the parameters you want to use to obtain most informative features. Before the feature extraction decide whether you need preprocessing including (among others) removal of non-informative instances.

   (a) Describe preprocessing steps if any. **Max. two sentences.**

   (b) Describe the parameters that you set for the vectorizer. Explain your reasoning? **Max. one sentence per parameter.**

   (c) How many features did you extract? Why? **Max. one sentence.**

   Answer (a) - Preprocessing:

   -

   Answer (b) - Feature computation:

   -

   Answer (c) - Number of features:

   -

2. *Clustering.* Using the features that you extracted implement a clustering method of your choice. Use an appropriate evaluation metric to evaluate the quality of your clustering result.

   (a) What is your clustering algorithm and why? **Max. two sentences.**

   (b) How many clusters did you extract? How did you decide on the number of clusters. **Max. one sentence.**

   (c) Which evaluation metric did you use to evaluate your results. What is your evaluation score? **Max. two sentences.**

   (d) Interpret your clusters, e.g., by looking into ten most important words in each cluster. **Max. one sentence per cluster.**

   Answer (a) - Clustering algorithm:

   •

   Answer (b) - Number of clusters:

   •

   Answer (c) - Evaluation:

   •

   Answer (d) - Interpretation:

   •

3. *Dimensionality Reduction for Visualization.* Perform dimensionality reduction with PCA on the features that you extracted previously. Use your clustering results and plot data points in 2D PCA space with clusters as colors for your data points.

   (a) Your plot.

   (b) Are clusters well separated in your plot? **Max. one sentence.**

   (c) Interpret the PCA dimensions that you used for visualization. **Max. one sentence per dimension.**

Answer (a) - Your plot:

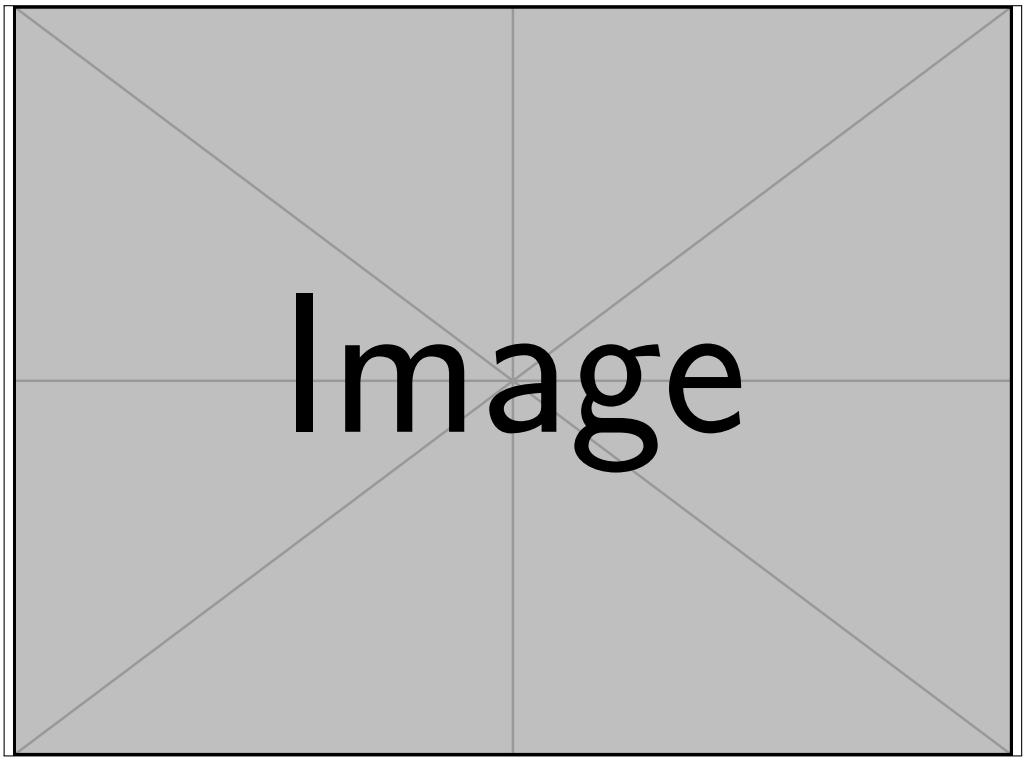

Figure 1: Cluster quality vs. number of clusters

Answer (b) - Cluster separation:

-

Answer (c) - Interpretation:

- PCA-1 ...
- PCA-2 ...

4. *Classification.* Given the dataset "king_rook_vs_king.csv" (available in TeachCenter), with data on chess endgames featuring the white king and a white rook against the black king, implement a classifier of your choice to predict whether the white will win. Each endgame is described by the rank and file positions of the white king, the white rook, and the black king (six features in total). The target variable is the depth of white win (a categorical variable with either draw or zero, one, ..., sixteen indicating that the white wins in that many moves). Transform the target variable to obtain the win depth levels as:

   - draw: 0
   - zero, one, two, three, four: 1
   - five, six, seven, eight: 2
   - nine, ten, eleven, twelve: 3
   - thirteen, fourteen, fifteen, sixteen: 4.

   Use this new variable as your classification target. Evaluate your classifier by a metric of your choice. If your model has hyperparameters cross-validate.

   (a) Describe preprocessing and feature transformations steps if you made any. **Max. two sentences.**

   (b) What is your model and why? **Max. two sentences.**

   (c) Describe your evaluation setup. **Max. one sentence.**

   (d) Describe hyperparameter optimization if any. Give the final values of hyperparameters. **Max. two sentences.**

   (e) Give your evaluation results as text or a table.

   Answer (a) - Preprocessing & feature transformations:

   -

   Answer (b) - Model choice:

   -

   Answer (c) - Evaluation setup:

   -

   Answer (d) - Hyperparameters:

   -

   Answer (e) - Results:

   -