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:

   - The text was converted to lowercase. Punctuation and stop words were eliminated, and lemmatization was applied to ensure that only informative and meaningful tokens were retained.

   Answer (b) - Feature computation:

   - max_df=0.85: To ignore words that appear in more than 85% of talks.
   - min_df=5: To ignore words that appear in less than 5 talks.
   - ngram_range=(1, 1): Only unigrams (single words) were considered as features during the feature extraction process, and bigrams or higher n-grams were excluded.
   - stop_words='english': To remove common English stopwords.

   Answer (c) - Number of features: 11008

   - The 11,008 features represent the unique words (unigrams) found in your dataset, and this number is due to the variety of words in the documents and the parameters used in the TF-IDF vectorization.

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:

- **K-Means Clustering** was chosen as it is a simple yet efficient algorithm, particularly effective for large datasets with well-separated clusters.
- Its ability to minimize the variance within each cluster made it a suitable option for this task, especially when the number of clusters was known or could be reasonably estimated.

Answer (b) - Number of clusters:

- A total of **16 clusters** were extracted based on the Elbow Method. This technique helps identify the optimal number of clusters by finding the point where the decrease in inertia begins to slow down.

Answer (c) - Evaluation:

- The Silhouette Score of 0.0123 indicates that the clusters are not well separated and likely have significant overlap. This suggests that the clustering may not be capturing distinct patterns in the data.

Answer (d) - Interpretation:

- Cluster 1: Focuses on agriculture and farming, including issues related to food production, fertilizers, and farmers' needs.
- Cluster 2: Discusses Schengen area enlargement and border issues involving Romania, Bulgaria, and Croatia.
- Cluster 3: Centers on the European energy crisis, covering gas prices, electricity, and renewable energy.
- Cluster 4: Highlights public health topics, especially child and youth mental health and cancer care.
- Cluster 5: Addresses gender equality and women's rights, including topics like abortion and gender-based violence.
- Cluster 6: Contains general political discourse, including addresses to the president and broad discussions on rights and opinions.
- Cluster 7: Relates to procedural parliamentary matters, such as statements, votes, and scheduling.
- Cluster 8: Covers EU foreign relations, particularly with neighboring countries like Moldova and broader European identity.
- Cluster 9: Focuses on internal documentation and procedural reports within the European Parliament.
- Cluster 10: Discusses rule of law, democracy, and press freedom, with attention to Hungary and human rights.
- Cluster 11: Concerns administrative actions such as appointments, objections, and parliamentary notifications.
- Cluster 12: Focuses on the Ukraine–Russia war, highlighting support for Ukraine and criticism of Putin.
- Cluster 13: Covers climate policy and EU member state coordination on social and environmental issues.
- Cluster 14: Involves parliamentary voting procedures, session management, and agenda items.

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:

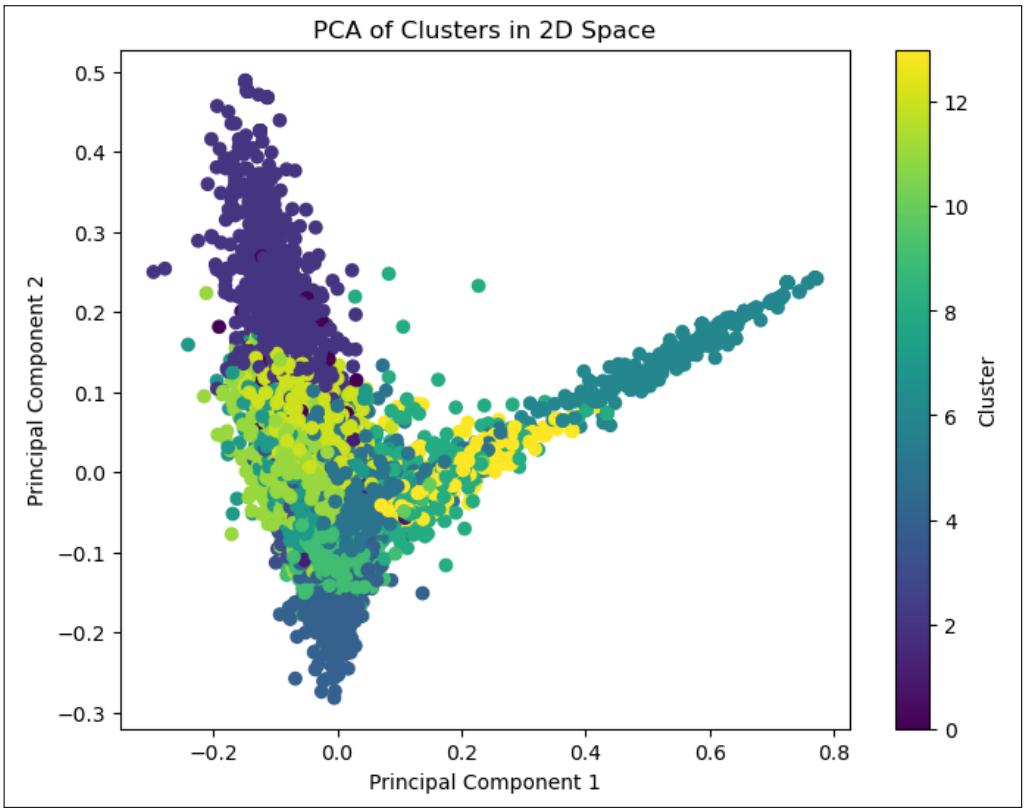

Figure 1: 2D PCA Projection of K-Means Clusters

Answer (b) - Cluster separation:

- No, the clusters appear to overlap in the plot, indicating that they are not well separated.

Answer (c) - Interpretation:

- PCA-1: "Parliamentary Procedure vs. Thematic Policy Content" — higher PC1 values correspond to debates focused on formal session management and logistics.
- PCA-2: "Policy Intensity on Energy/Climate/Economic Issues" — higher PC2 values reflect speeches focused on urgent policy concerns like the energy crisis and environmental sustainability.

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:

- One-hot encoding was applied to categorical columns such as white_king_file, white_rook_file, and black_king_file. The target column, white_depth_of_win, was transformed into categorical levels, where "draw" was assigned a value of 0, and the remaining outcomes were grouped into four levels based on the depth of the win (from 1 to 4).
- Standard scaling was applied to both the training and test sets to ensure the features were on a comparable scale.

Answer (b) - Model choice:

- **The Decision Tree Classifier** was used because it is easy to interpret, handles both categorical and numerical features effectively, and can capture the positional logic of chess endgames. Its ability to model non-linear relationships makes it well-suited for learning strategic patterns from piece placements.

Answer (c) - Evaluation setup:

- To ensure consistent performance across the dataset, the model was evaluated using 5-fold cross-validation with **accuracy** as the primary metric, and further analyzed through a **confusion matrix** and **classification report** to understand how well it performed on each class.

Answer (d) - Hyperparameters:

- GridSearchCV was used to tune the Decision Tree Classifier for optimal performance.
- The best parameters found were: **criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features=None**, and **max_leaf_nodes=None**.

Answer (e) - Results:

Table 1: Confusion Matrix for Win Depth Level Classification

	Pred 0	**Pred 1**	**Pred 2**	**Pred 3**	**Pred 4**
Actual 0	408	6	50	50	35
Actual 1	14	99	15	2	0
Actual 2	35	9	540	50	2
Actual 3	46	3	48	1894	81
Actual 4	53	0	5	74	2093

Table 2: Classification Metrics by Class

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.73	0.74	0.74	549
1	0.85	0.76	0.80	130
2	0.82	0.85	0.83	636
3	0.91	0.91	0.91	2072
4	0.95	0.94	0.94	2225
Accuracy	0.90 (Total samples: 5612)			
Macro Avg	0.85	0.84	0.85	5612
Weighted Avg	0.90	0.90	0.90	5612