### Question 1: Feature Engineering

**Answer (a) - Preprocessing:**

* Removed talks with fewer than 10 words to ensure meaningful content.
* Converted text to lowercase, removed punctuation and stop words, and lemmatized words for consistency.

**Answer (b) - Feature computation:**

* max\_df=0.8: Ignores words in over 80% of documents to exclude overly common terms.
* min\_df=5: Ignores words in fewer than 5 documents to filter out rare, uninformative terms.
* ngram\_range=(1,1): Uses unigrams to capture individual word importance initially.
* max\_features=1000: Limits features to 1000 to manage dimensionality and focus on key terms.

**Answer (c) - Number of features:**

* Extracted 1000 features to balance computational efficiency with capturing significant topics.

Question 2: Clustering

Answer (a) - Clustering algorithm:

Chose K-Means for its efficiency and compatibility with TF-IDF features.

It’s widely used for text clustering due to its simplicity and scalability.

Answer (b) - Number of clusters:

Selected 10 clusters using the elbow method, identifying a bend at k=10.

Answer (c) - Evaluation:

Used silhouette score to assess cluster quality, achieving 0.25.

This indicates moderate separation, suggesting reasonably defined clusters.

Answer (d) - Interpretation:

Cluster 1: "europe", "union", "policy" - Focuses on EU governance.

Cluster 2: "democracy", "freedom", "rights" - Emphasizes democratic values.

Cluster 3: "pandemic", "health", "crisis" - Relates to COVID-19 responses.

Cluster 4: "solidarity", "social", "support" - Highlights social unity themes.

Cluster 5: "climate", "energy", "sustainability" - Centers on environmental issues.

Cluster 6: "war", "ukraine", "russia" - Discusses Ukraine conflict.

Cluster 7: "economy", "budget", "fund" - Pertains to financial matters.

Cluster 8: "migration", "asylum", "border" - Addresses migration policies.

Cluster 9: "justice", "law", "rule" - Focuses on legal frameworks.

Cluster 10: "david", "sassoli", "president" - Reflects memorial tributes.

Question 3: Dimensionality Reduction for Visualization

Answer (a) - Your plot:

Answer (b) - Cluster separation:

Clusters show some separation but overlap, indicating moderate distinctiveness.

Answer (c) - Interpretation:

PCA-1: Driven by economic and policy terms like "economy" and "union."

PCA-2: Influenced by social and humanitarian terms like "solidarity" and "rights."

Question 4: Classification

Answer (a) - Preprocessing & feature transformations:

Mapped file letters to numbers (a=0, b=1, ..., h=7) and kept rank numbers as is.

Transformed target variable into levels: draw=0, zero-four=1, five-eight=2, nine-twelve=3, thirteen-sixteen=4.

Answer (b) - Model choice:

Selected Random Forest for its robustness with categorical data post-encoding.

It handles mixed data types effectively and provides strong predictive performance.

Answer (c) - Evaluation setup:

Used 5-fold cross-validation with accuracy to assess model performance.

Answer (d) - Hyperparameters:

Tuned n\_estimators and max\_depth via grid search, yielding 100 and 10, respectively.

These values optimized accuracy while preventing overfitting.

Answer (e) - Results:

Achieved 85% accuracy, with a confusion matrix showing balanced class performance.

