Project 2: Exploring Textual Data Analysis Using SpaCy

Form groups of either two individuals or one individual based on preference. If you create a group, send me an email and cc your colleague.

1. Data Preprocessing with SpaCy:

- Utilize SpaCy for text preprocessing tasks such as tokenization, lemmatization, and removing stop words. Ensure the dataset is cleaned and prepared for modeling.

2. Model Selection and Group Formation:

- Each group will then select **four models** for classification or clustering tasks or other tasks
- Each individual will select **two models** for classification or clustering tasks or other tasks

3. Model Training and Hyperparameter Tuning:

- Train selected models on preprocessed data. Each model should be trained with two hyperparameters, and each hyperparameter should be tested with two different values to explore various model configurations.

- For classification tasks, ensure at least **one model utilizes SpaCy** for classification. Other models can include algorithms like SVM, Random Forest, or Neural Networks or ….
- For clustering tasks, consider models like K-means, DBSCAN, or hierarchical clustering or ….

4. Performance Evaluation:

- Select appropriate performance metrics based on the nature of the dataset and the task (classification or clustering). Common metrics include accuracy, precision, recall, F1-score for classification, and silhouette score or Davies-Bouldin index for clustering.
- Evaluate the performance of each model with different hyperparameter configurations to identify the most effective model settings.

5. Analysis and Conclusion:

- Analyze the results obtained from various models and hyperparameter configurations.
- Discuss the impact of SpaCy in comparison to other models in terms of performance and computational efficiency.
- Draw conclusions on the suitability of different models and hyperparameter settings for the given dataset and task.