Assessment 3: Clustering News Headlines

Overview

In this assignment, you will apply clustering techniques to analyze a dataset of 19,685 news headlines. These headlines have been preprocessed and transformed into a 1,000-dimensional feature vector using the **TF-IDF (Term Frequency-Inverse Document Frequency)** method. While the details of TF-IDF are beyond the scope of this course, you can explore the following resources for additional reading:

- Introduction to TF-IDF
- Scikit-learn TF-IDF Documentation

Your task is to uncover patterns in the data using **K-Means Clustering**. You will also compare your results briefly with **Hierarchical Clustering** on a smaller subset of the data.

Dataset

You have been provided the following files in this link:

1. `headlines.csv`: A CSV file containing 19,685 news headlines.
2. `tfidf_features.npy`: A NumPy file containing the corresponding TF-IDF feature matrix with shape `(19,685, 1000)`.

Tasks

1. Load and Explore the Dataset

- Load the dataset and the feature matrix into Python.
- Check the shapes of the data to confirm successful loading.

2. Perform K-Means Clustering

- Perform K-Means clustering on the TF-IDF feature matrix.
- Identify the optimal number of clusters (k) and run K-Means with this k.
- Assign headlines to clusters and analyze the contents of at least three clusters:
  - Select three clusters and display 10 representative headlines for each.

3. Perform Hierarchical Clustering (Smaller Subset)

- Use Agglomerative Hierarchical Clustering on a **random subset of 1,000 headlines** from the dataset.
- Visualize the clustering structure using a dendrogram.
- Assign these 1,000 headlines to clusters based on your analysis of the dendrogram.

4. Compare and Summarize

- Compare the themes in the clusters produced by K-Means and Hierarchical Clustering.
- Write a short summary of your observations in your report.

Optional Bonus Task (Not Graded)

- Explore clustering after **dimensionality reduction** using Principal Component Analysis (PCA). Reduce the TF-IDF matrix to 50 dimensions and apply K-Means again. Compare the results to the original clustering.

Additional Notes

- If you encounter memory issues, you may select a subset of at least **5,000 sentences** for the entire assignment.
- You need **not** implement either of the clustering algorithms and can use existing implementations/API calls

Deliverables

1. **Jupyter Notebook** (`.ipynb`):
    - Include all your code, visualizations, and brief explanations.
    - Ensure it runs without errors.
2. **Short Report** (`.pdf` or `.docx`):
    - Summarize your findings, including:
        - The optimal number of clusters for K-Means.
        - Themes observed in two K-Means clusters.
        - Insights from the dendrogram and Hierarchical Clustering.

Submission Guidelines

- Submit your notebook and report via the LMS by **30th November, EOD**.
- Ensure the notebook and report are clear and well-structured.

Evaluation Criteria

1. **Clarity and Completeness (60%)**:
    - Are the tasks completed clearly and correctly?
    - Are the themes in the clusters well-explained?
2. **Code Quality (40%)**:
    - Is the code clean, efficient, and well-commented?

Resources

1. K-Means Clustering
2. Hierarchical Clustering

Good luck, and we look forward to your insights!