Assignment 2: Extracting Topics from the Documents

Objective

This assignment aims to help you understand the fundamentals of topic modeling, preprocessing text for topic modeling, and evaluating the generated topics.

Instructions

Complete the tasks below. Each task specifies the marks assigned. Submit your code, outputs, and a brief explanation for each step.

Dataset: text_docs

Tasks

Task 1: Data Exploration (3 Marks)

Instructions:

1. Load the dataset **text_docs** provided.
2. Print the total number of rows and any relevant statistics, such as the number of unique documents.
3. Identify any preprocessing steps that might be required.

Task 2: Generate Topics Using LDA (7 Marks)

Instructions:

1. Prepare the data for LDA by creating a document-term matrix using a library like gensim or sklearn.
2. Apply Latent Dirichlet Allocation (LDA) to extract topics from the dataset. Choose a suitable number of topics (e.g., 5).
3. Display the top 5 words for each topic generated by the model.

Submission Guidelines

- Submit your Python code and output in a single Jupyter Notebook or script file.
- Include a brief explanation for each task in markdown cells or comments.

Note: Ensure that all code is well-commented, and outputs are clearly displayed.