Report on NLP Assignment 2 - Task 4

Manish Kanuri
NUID: 002315456

Objective

The goal of this assignment is to preprocess and analyze text data from the **20 Newsgroups dataset** using various NLP techniques.

Dataset

- The dataset includes text from 10 selected categories:
  - `alt.atheism`
  - `comp.graphics`
  - `comp.os.ms-windows.misc`
  - `comp.sys.ibm.pc.hardware`
  - `rec.autos`
  - `rec.sport.baseball`
  - `sci.crypt`
  - `sci.space`
  - `talk.politics.guns`
  - `talk.religion.misc`
- Headers, footers, and quoted text were removed to focus on meaningful content.

Preprocessing Steps

- Converted text to lowercase
- Removed non-alphanumeric characters
- Removed numerical digits

Implementation Highlights

- **Text Cleaning**: A function was defined to preprocess the dataset.
- **Query Matching**: A set of predefined queries was used to retrieve relevant documents.
- **Evaluation Metrics**: Precision, were calculated to assess query performance.

Results

- The dataset was successfully loaded and preprocessed.
- The system retrieved and ranked documents based on query relevance.
- Performance metrics indicated the effectiveness of the approach.

Conclusion

This assignment demonstrated key text preprocessing techniques and retrieval evaluation. Further improvements could involve implementing advanced ranking models or exploring deep learning-based approaches for text classification.