Week 8: Deliverables

Group Name: NLP_Task_Force – Document Classification

Name: Konstantinos Soufleros

Email: soufleros.kostas@gmail.com

Country: Serbia

Company: Data Glacier

Specialization: NLP

Github: https://github.com/kostas696/DG_Intern/tree/main/week8

Internship Batch: LISUM30

Date: 26/03/2024

PROBLEM DESCRIPTION

The problem revolves around analyzing a dataset containing newsgroup documents categorized into various topics. The primary objective is to gain insights into the content of these documents, understand the prevalent themes within each newsgroup, and identify any patterns or trends present in the data.

DATA UNDERSTANDING

The dataset consists of documents from 20 different newsgroups, covering a wide range of topics such as sports, religion, politics, technology, and more. Each document is labeled with its corresponding newsgroup category, providing a structured format for analysis.

TYPE OF DATA

The data is structured and tabular, organized into a pandas DataFrame with two columns: 'Newsgroup' and 'Content'. The 'Newsgroup' column contains categorical labels indicating the category of each document, while the 'Content' column contains the textual content of the documents.

DATA PROBLEMS

- No missing values: The dataset does not contain any missing values in either the 'Newsgroup' or 'Content' columns, as confirmed by checking for null values.
- Document length variation: The length of documents varies across different newsgroups, with some containing longer texts compared to others. This variation in document length might affect certain analyses or models.
- Presence of special characters, numbers, and stopwords: The textual content contains special characters, numbers, and stopwords that may not contribute meaningfully to the analysis. These elements need to be removed or filtered out to focus on relevant textual features.
- Need for preprocessing: The textual content requires preprocessing to standardize the format, remove unnecessary elements, and prepare it for further analysis or modeling tasks.

APPROACHES TO ADDRESS DATA PROBLEMS

Text Preprocessing: A preprocessing function is applied to clean and standardize the textual content. This function involves:

- Removal of metadata headers, emails, numbers, and 'GMT' mentions.
- Tokenization to split the text into individual words.
- Lowercasing to ensure consistency in word case.
- Removal of punctuation, non-alphabetic characters, single characters, and stopwords.
- Lemmatization to reduce words to their base or dictionary form.

Word clouds and histograms are generated to visualize the distribution of words and document lengths across different newsgroups. These visualizations help in identifying common themes, prevalent words, and document length patterns within each category.

Average document lengths per newsgroup are calculated to understand the variation in document sizes across different categories.

CONCLUSION

By understanding the nature of the data, identifying potential issues, and applying appropriate preprocessing and analysis techniques, we aim to gain valuable insights into the content and structure of the newsgroup documents, enabling better understanding and interpretation of the dataset.