Week 9: 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/week9

Internship Batch: LISUM30

Date: 02/04/2024

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 'Original_Content'. The 'Newsgroup' column contains categorical labels indicating the category of each document, while the ''Original_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

Previous week in our Original Method we preprocessed text data using traditional methods, including metadata removal, tokenization, and lemmatization, visualized document length distribution and common words in original content. This Week using the SpaCy Method we:

- Utilized SpaCy model for preprocessing, including tokenization and lemmatization.
- Applied SpaCy preprocessing to 'Original_Content' column.
- Visualized document length distribution and common words in SpaCy preprocessed content.
- Recalculated average document length per newsgroup.
- Compared document length distribution and common words between original and SpaCy preprocessed content.
- Assessed differences in preprocessing impact on document lengths and common words.

Results:

Both methods effectively cleaned text data, but SpaCy offered streamlined preprocessing. Document length and common word distributions varied slightly between methods. SpaCy preprocessing showed potential improvements in tokenization and lemmatization.

CONCLUSION

SpaCy-based text preprocessing offers a more efficient and comprehensive approach compared to the original method. By leveraging advanced NLP capabilities, SpaCy improves the accuracy and consistency of preprocessing tasks, leading to better-quality textual data for subsequent analysis and modeling. The comparison highlights the benefits of adopting SpaCy for text preprocessing in data science workflows.