Williane Yarro

Assignement Name : NewsBot_ITAI2373-New-Mdterm

Instructor Name: Dr. Patricia Mcc Manus

Assignement/Lab : NewsBot Intelligence System

Introduction:

This project consolidates all the Natural Language Processing (NLP) techniques discussed in Modules 1 through 8 into a single cohesive system. The objective was to create a comprehensive text processing pipeline—from initial raw input to valuable insights—while implementing various machine learning models for classification and analysis. I started by uploading and configuring the Kaggle dataset, resolving path issues and filename discrepancies that arose from multiple uploads. I faced several technical challenges, including permission errors when copying `kaggle.json`, the need to manually rename the file (for instance, `kaggle (3).json`), and utilizing the `-o` flag to unzip files forcefully. Furthermore, I had to meticulously manage sampling, address missing values, and ensure consistent column alignment. These obstacles provided me with valuable experience in troubleshooting real-world problems in NLP workflows. In the end, this project enhanced my capability to extract structured information from text, assess model performance, and effectively communicate results to both technical and business stakeholders.

Activities Conducted and Assignment Feedback

To fulfill the learning goals of this project, I undertook a variety of tasks independently, utilizing NLP techniques discussed in Modules 1–8. The focus of the project was on classifying news headlines with a dataset sourced from Kaggle, necessitating the creation of a comprehensive NLP pipeline and addressing numerous technical obstacles.

◇ Activities Conducted

Dataset Selection and Configuration

I selected the News Category Dataset from Kaggle because of its relevance to real-world scenarios and its variety. I set up the Kaggle API in Google Colab and uploaded the kaggle.json token for authentication. Due to multiple uploads, the token filename was altered (e.g., kaggle (3).json), which resulted in errors. I manually rectified the path using:

- !cp 'kaggle (3).json' ~/.kaggle/kaggle.json
- -o

File Download and Extraction

Upon downloading the dataset via the API, I encountered prompts requesting confirmation to overwrite files while extracting. This issue was addressed by utilizing the -o flag:

- !unzip -o news-category-dataset.zip

Data Import and Preparation

I utilized the json module to load the dataset and transformed it into a Pandas DataFrame. To prevent memory overload in Colab, I generated a sample of 2,000 rows using:

- df_sample = df.sample(n=2000, random_state=42)

Preprocessing Steps

The headline was identified as the text field, while the category was designated as the label.

Missing values were checked for and subsequently removed.

Class distribution was verified to confirm sufficient representation.

Columns were renamed for improved clarity.

df_final = df_clean.rename(columns={'headline': 'content', 'category': 'category'})

Error Management and Troubleshooting

During the process, I faced and addressed several issues, including:

FileNotFoundError caused by incorrect filenames.

Memory constraints in Colab resolved through downsampling.

During the process, I faced and addressed several issues, including:

FileNotFoundError caused by incorrect filenames.

🎓 Insights Gained

Kaggle API Setup in Google Colab

I acquired knowledge on how to securely set up the Kaggle API within a Colab notebook environment. This encompassed managing file permissions and rectifying path issues that arose from multiple uploads (e.g., kaggle (3).json).

Practical Data Management

Engaging with the BBC News dataset highlighted the necessity of exploring, examining, and restructuring real-world textual data prior to modeling. This process involved parsing extensive files, pinpointing essential columns, and ensuring uniform formatting.

Troubleshooting and Command Line Skills

I obtained hands-on experience in resolving file and unzip errors by utilizing flags such as -o to enforce overwrites, as well as employing shell commands within Colab to organize folders and files.

Foundation for NLP Model Implementation

While not executed in the notebook, the essential groundwork was established for implementing NLP models. Preparations were made for tasks including text vectorization, label encoding, and model training.

By the conclusion of the project, I successfully curated a clean, balanced dataset that fulfills all criteria and is prepared for the modeling stage. Additionally, I gained a more profound understanding of both the technical and practical elements involved in constructing a comprehensive NLP pipeline.

Works Cited

Kaggle. *BBC News Classification Dataset*. Kaggle, www.kaggle.com/competitions/learn-ai-bbc. Accessed 28 July 2025.

Pedregosa, Fabian, et al. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830. www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Hunter, John D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90–95. doi:10.1109/MCSE.2007.55.

Waskom, Michael, et al. *Seaborn: Statistical Data Visualization*. seaborn.pydata.org, Accessed 28 July 2025.

Python Software Foundation. *Python Language Reference, Version 3.10*. docs.python.org, Accessed 28 July 2025.

The Pandas Development Team. *Pandas Documentation*. pandas.pydata.org, Accessed 28 July 2025.

Google. *Colaboratory Documentation*. research.google.com/colaboratory, Accessed 28 July 2025.