**# News Summarization**

This project, is part of the NewsNuggets application. This section presents the results of implementing three distinct extractive summarization techniques on the BBC News Dataset. The objective is to evaluate and compare the performance of Word Frequency-based Extractive Summarization, TF-IDF-based Summarization, and TextRank Summarization. Each technique represents a unique approach to the challenge of text summarization, which is critical in managing the vast amount of information generated daily.

**## Dataset Used:**

The BBC Dataset consists of 2,225 documents collected from the BBC News website, corresponding to stories published in five topical areas during 2004-2005. It has a News document with its summary

Dataset Source:<http://mlg.ucd.ie/datasets/bbc.html>

**## Project Structure:**

- **DataLoader/:** Contains data\_loader.py for loading the dataset.

- **DataPreprocessing/:** Contains data\_preprocessing.py for text preprocessing including removal of stopwords, punctuations, tokenization, and more.

- **Modeling/:** Contains TextRank, TfIdf, and WordFrequency for training news summarization models.

- Evaluation/: Contains evaluate.py which is the code for evaluating the summarization models

- **Results/:** Storage for all models and their evaluation results.

- **logs/:** Contains logs of the processes.

**## Prerequisites:**

- **Python** 3.x

- **Libraries**: pandas, numpy, scikit-learn, nltk, scipy, rouge, nx

- Ensure **NLTK** resources (punkt, stopwords, wordnet) are downloaded as the code will attempt to do so.

**## Setup:**

1. Unzip the 'News\_Summarization.zip' file to your local machine.

2. Navigate to the unzipped directory.

3. Set up a virtual environment:

4. Install required Python packages using the requirements.txt file :

pip install -r requirements.txt

**## Configuration:**

Adjust the logging and file path settings in main.py according to your local environment.

Running the Code:

To run the project, execute the main.py script:

**python main.py**

This script performs the following actions:

- Initializes logging.

- Loads the dataset from the specified path.

- Preprocesses the text data.

- Trains different models for news summarization.

- Evaluate models and save the outputs and pickle file in the Results directory.

**Estimated Running Time:**

- The complete execution of the script may take approximately 10 minutes, depending on your machine's specifications and the size of the dataset.

**## Outputs:**

The system will generate:

- Trained model files (\*.pkl).

- Model evaluation summaries and detailed reports.

**## Logging:**

Logs are written to application.log in the logs directory, capturing detailed events and errors throughout the execution process.

**## Authors:**

- [ANITHA BALACHANDRAN, MADHURA BHATSOORI, SOWJANYA PAMULAPATI,

SOWMYA KURUBA, SWATHI RAMESH ]

**Note:**

While executing main.py, you may encounter "UserWarning" messages due to the Blue score n-gram selection methods. These warnings are not necessarily indicative of an issue with your code or the optimization process itself.

**Team Contribution:**

A table with text on it

Description automatically generated