Blackcoffer NLP and Data Extraction Project

Python-based Text Analysis and Data Extraction

Objective

- To extract textual data from articles and perform text analysis using Python.

Data Extraction

- Extract article text from given URLs, ensuring only article content is retrieved.

- Save extracted text using URL_ID as file names.

- Python libraries: BeautifulSoup, Selenium, Scrapy.

Text Analysis Variables

- For each extracted article, compute the following variables:

- - Positive Score

- - Negative Score

- - Polarity Score

- - Subjectivity Score

- - Average Sentence Length

- - Word Count and others

Python Code Example: Web Scraping

```python
from bs4 import BeautifulSoup
import requests

url = 'http://example.com'
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')
article = soup.find('div', class_='article').get_text()
```

Saving Results

- After analysis, the results are saved in the 'Output Data Structure.xlsx' file in CSV/Excel format.

- Results include sentiment scores, readability metrics, and word/sentence-level statistics.