Data Science Project Assessment

Objective

The goal of this assignment was to programmatically scrape data from news articles using their URLs, extract named entities (persons or organizations) mentioned in the articles, and classify the sentiment of the articles as positive, negative, or neutral. This report outlines the approach taken, challenges faced, and reflections on the results.

Approach

1. Data Scraping

Using Python, the BeautifulSoup library was employed to scrape the HTML content of news articles from their respective URLs. The steps included:

- Fetching the webpage content using the requests library.
- Parsing the HTML content with BeautifulSoup to extract the article text, avoiding unwanted sections such as advertisements or navigation links.
- Preprocessing the extracted text to remove noise like stopwords, punctuation, and special characters.

2. Entity Extraction

To identify named entities such as persons and organizations, the **spaCy** NLP library was utilized:

- The nlp pipeline from spaCy was applied to the cleaned text to detect and classify entities.
- Extracted entities were categorized into types like PERSON and ORG for further analysis.
- Results were stored in a structured format for review and interpretation.

3. Sentiment Analysis

Sentiment classification was performed using the **TextBlob** library:

- Sentiment polarity scores (ranging from -1 to 1) were calculated for the article text.
- Thresholds were set: scores > 0 indicated positive sentiment, scores < 0 indicated negative sentiment, and scores close to 0 were deemed neutral.

Challenges Faced

1. **Variability in Webpage Structures**: Different news websites have unique layouts, making it challenging to standardize text extraction. Additional fine-tuning was required to locate the correct HTML tags.
2. **Entity Extraction Limitations**:
    - SpaCy occasionally missed entities or incorrectly classified them due to ambiguous contexts.
    - Noise in the text (e.g., incomplete sentences or poorly formatted data) impacted accuracy.
3. **Sentiment Ambiguity**: Articles often contained mixed sentiments, making it difficult to classify them definitively. For instance, a generally positive article could contain negative remarks that skew the sentiment score.