

Data Science Project Assessment

Objective: The objective of this assignment is to demonstrate your ability to programmatically scrape data from news articles using their URLs, extract entities (persons or organizations) mentioned in the articles, and classify the sentiment of the articles as positive, negative, or neutral.

Tasks: 1. Data Scraping:

- Write a Python script that takes a news article URL as input and retrieves the HTML content of the article.
- Bonus Point: Create UI which takes input and shows result.

2. Entity Extraction:

- Implement Named Entity Recognition (NER) to identify and extract entities (persons or organizations) mentioned in the article.
- Filter and store entities that are classified as 'PERSON' or 'ORG'.
- Bonus Point: Train your own model.

3. Sentiment Analysis:

- Choose a method to analyze the sentiment of the article. You can use pretrained.
 - Classify the sentiment of the article as positive, negative, or neutral based on the extracted main text.

4. Storage:

- Store the request response in the database in proper schema.

5. Deployment:

- Deploy the full project on docker local environment (Bonus Point)

6. Documentation and Reporting:

- Provide clear documentation within your Python script explaining each function and its purpose.
- Write a brief report (1-2 pages) summarizing your approach, challenges faced (if any), and reflections on the accuracy of entity extraction and sentiment analysis.

7. Maintain Code in Repository:

- After completion of the project push your code into GitHub repository.

Deliverables:

- Python script(s) containing the implementation of the scraping, entity extraction, and sentiment analysis functionalities.
- Test the script with URLs to demonstrate the correct extraction of entities and sentiment classification.
- Report document summarizing your approach and findings.

This assignment will assess your ability to integrate multiple aspects of data processing (web scraping, NER, sentiment analysis, deployment, database) into a cohesive application, as well as your skills in documenting and reporting on your work.

Note* If you are qualified for the interview round, please be ready with running the project otherwise you will be disqualified.