UNIVERSIDADE
BEIRA INTERIOR

FACULDADE
ENGENHARIA
Departamento de
Informática

PTNewsAnalyzer: Leveraging NLP for News Article Analysis

Proposta de Projeto. Lic. Eng. Informática

Orientador: Ricardo Campos (ricardo.campos@ubi.pt)

Coorientador: João Canavilhas (jc@ubi.pt)

Objectives

Imagine the challenge of sifting through a multitude of Portuguese news articles and online content without the aid of a robust system for information extraction?

Information extraction, a fundamental natural language processing (NLP) technique, involves automatically identifying and extracting the most relevant information from a given text. This process serves, among other things, to provide a concise summary of the text's content and the primary topics discussed within it. Such capabilities are invaluable for various downstream tasks, including text analysis, social media monitoring, text summarization, text tagging, and information retrieval, to name but a few.

In this project, the student will embark on the task of collecting news articles from Portuguese web sources over time and implementing NLP techniques to extract pertinent information from the texts. Examples of this are relevant keywords, topics discussed overtime, etc. The extracted information will then be utilized by journalists and researchers of the LabCom R&D unit of the University of Beira Interior to produce weekly reports covering a range of topics, including politics, sports, and more.

Workplan

T1: Project setup and planning (2 weeks)

T2: Development of Python scripts to collect news articles from Portuguese web sources on a regular basis. This will involve setting up web scrapers or utilizing existing APIs to retrieve articles from various news websites (e.g., Jornal Público, Observador, etc) (3 weeks)

T3: Implementation of methods for extracting relevant information from the collected news articles. This may include identifying key topics, named entities, relevant keywords, sentiment analysis, and other relevant NLP tasks (3 weeks)

T4: Creation of a Docker image to integrate the various components (0.5 weeks)

T5: Implementation of a storage and search system using a noslq database (e.g., elasticsearch or redis) and information retrieval algorithms to retrieve the most relevant information (0.5 weeks)

T6: Development of a website (using flask or streamlit) to provide easy access to parts of the information extraction system. This will enable journalists and other users to access extracted information for further analysis and reporting. (4 weeks)

T7: Report writing and presentation (2 weeks)

Table 1 presents the distribution of tasks for each of the 15 weeks.

Table 1: Chronology of tasks (T) per week (S).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
T1	▨	▨													
T2			▨	▨	▨										
T3						▨	▨	▨							
T4									▨						
T5									▨						
T6										▨	▨	▨	▨		
T7														▨	▨

Technical and Academic Prerequisites

- **Proficiency in Python**: Strong programming skills in Python (or willingness to learn)
- **Natural Language Processing (NLP) Knowledge**: A solid understanding of NLP concepts, including text preprocessing, feature extraction, and information extraction algorithms.
- **Web Development Skills**: Proficiency in web development technologies (e.g., Flask or Streamlit).
- **nosql databases**: experience with nosql databases (elasticsearch, redis) or willingness to learn. **Virtualization**: familiarity with creating Docker images.

Expected Results

- Python script for data acquisition, storage and retrieval
- Compiled dataset
- Responsive website (code to be made available on the student's github) and available online
- Report