ASEC 2024 Conference

The 5th International Electronic Conference on Applied Sciences

04–06 December 2024 | Online

MDPI

Acquiring News Texts about Public Security for the construction of Corpora in Portuguese

Matheus Nascimento, Vagner Silva, Gabriel Souza, Kauã Lima, Jean Turet, Victor Diogho Heuer de Carvalho, Thyago Nepumoceno

Federal University of Alagoas

Group of Engineering in Decision-Making and Artificial Intelligence

INTRODUCTION & AIM

The acquisition of texts for the purpose of composing corpora in specific domains from sources on the social web is a process that requires analyzing the structures of websites where the texts are published. This involves searching for specific fields to guide the access of responsible agents, known as scrapers. With these texts in hand, performing more refined analyses focused on tasks such as named entity recognition, text summarization, sentiment mining, and associated classifications (e.g., opinion polarities) becomes possible. This article aims to demonstrate the process of acquiring news texts in the domain of public safety in Brazil to build corpora in the Portuguese language. Since Portuguese still lacks dedicated corpora on this topic, scraping agents were developed for three initial news sources in the Northeast region.

The constituted corpus enabled the execution of multiple preliminary analyses, including the identification of crime patterns, sentiment analysis in public security reports, and the mapping of risk areas. These analyses provided valuable information that can support the formulation of public policies and the development of more effective security strategies and decision-making.

METHOD

The methodology used considers the application of sophisticated news scraping techniques to build data collection algorithms on public safety in the Brazilian Northeast. News outlets were used as the collection source, and these tags were applied to specific public safety to obtain a more assertive collection.

Therefore, the Corpora development process was divided into 4 pillars: Algorithm Construction, Data Collection and Data Storage.

1. Algorithm Construction:

The creation of the algorithms considered the use of specific frameworks for collection, as well as the filtering of HTML from the news channels used for scraping. Using techniques such as **Scrapy**, **Selenium** and **Beautiful Soup.**

2. Data Collection:

Application of the algorithm for mass collection of texts with contingent information on public safety in the region.

3. Data Storage:

Deposit of the websites found in a cloud system from the accumulation of information in Json files.

4. Data Analysis:

Execution of sentiment analysis and Machine Learning algorithms to recognize crime patterns and map risk areas.

RESULTS & DISCUSSION

Decision Making

The use of the resulting corpora allowed the identification of crime patterns and the detailed analysis of risk areas in the Brazilian Northeast. This information was integrated into decision support models, providing support for strategic actions by security forces. Through data visualization and interpretation, managers will be able to prioritize resources and direct efforts more effectively, possibly impacting the reduction of crime rates in critical regions.

Sentiment Analysis

The sentiment analysis applied to the collected news revealed predominant perceptions and emotions in reports on public security. Text processing indicated trends, such as the level of concern in society regarding specific crimes and reactions to government measures. This information will help to better understand the psychological and social impact of violence, enabling a more humane and sensitive approach to dealing with the topic.

Shortage of Corpora in Portuguese

The project highlighted the lack of corpora dedicated to the domain of public security in the Portuguese language. The collection and organization of the texts filled a significant gap, establishing a valuable resource for the academic community and researchers. The corpora will now serve as a basis for future linguistic analyses and the development of Natural Language Processing (NLP) models adapted to the Brazilian reality.

Impact on Public Policies

The data analyzed will contribute to the formulation of more assertive and evidence-based public policies. The insights obtained have enabled the planning of initiatives aimed at preventing crimes and protecting vulnerable populations. In addition, the ability to map risk areas and predict criminal trends has generated a more proactive and efficient approach in the development of public safety strategies.

CONCLUSION

In addition to filling the gap in linguistic resources in the field of public safety, this project highlights the transformative potential of the use of technology and artificial intelligence in addressing complex social challenges. The results obtained demonstrate that initiatives like this can be replicated and expanded, contributing to the construction of a safer and more equitable environment. At the same time, they reinforce the importance of investing in interdisciplinary projects that integrate data science, computational linguistics, and public governance.

FUTURE WORK / REFERENCES

- Beautiful Soup(2024): https://beautiful-soup-4.readthedocs.io/en/latest/
- Selenium(2024):https://selenium-python.readthedocs.io/
- Scrapy(2023): https://scrapy.org/
- Sentiment analysis(2023): https://doi.org/10.1145/3664647.3689173