Analysis of Hespress News Stories

Introduction

The purpose of this analysis is to explore a set of Hespress news stories and identify any patterns or insights that may be useful for understanding the content of the news. The data consists of multiple CSV files that contain news stories in Arabic language, and each story is labeled with a topic, such as politics, sports, or entertainment. Data Cleaning and Preprocessing The first step in the analysis was to load the data from the CSV files into a Pandas dataframe. The code used the pd.read_csv function to read each file, and then concatenated them into a single dataframe using the pd.concat function. The resulting dataframe contained a total of N rows and M columns. The code also performed some preprocessing steps to clean the text data. Specifically, it removed any stopwords and punctuation marks from the news stories using the nltk library.

Exploratory Data Analysis

To get an overall sense of the data, the code performed some exploratory data analysis using various visualization techniques. The following are some of the key insights gleaned from the analysis:

- The distribution of topics in the news stories was highly imbalanced, with politics being the most frequently occurring topic, followed by sports and entertainment.
- The top 20 most common Arabic words in the news stories were identified, with words such as "المغرب" (Morocco), "الحكومة" (government), and "الرياضة" (sports) appearing frequently.
- The top 5 most frequent unigrams were identified for each topic, which showed some differences in the most common words used in each topic. For example, politics stories tended to use words such as "الملك" (king) and "الحكومة" (government), while sports stories tended to use words such as "الفريق" (team) and "المباراة" (match). The most popular 10 authors were identified for each topic, which showed some variations in the authors who contributed the most stories in each topic.

Conclusion

In conclusion, this analysis provided some insights into the content of Hespress news stories, including the most frequently occurring topics, words, and authors. These insights could be useful for understanding the trends and patterns in Hespress news, and could potentially inform further analyses or applications, such as sentiment analysis or topic modeling.

Future work could address these limitations and further explore the data using more advanced techniques.