Gufeng Shen
11/16/2023

Final Project Report

Analysis of Technology News Articles

I. Abstract

This report outlines a project undertaken to collect, process, analyze, and visualize data from technology news articles. Utilizing Python for data scraping, cleaning, and analysis, this study aims to uncover patterns and insights from the current technology news landscape.

II. Introduction

The rapid evolution of technology is a subject of significant interest in modern society. Analyzing technology news articles can provide insights into emerging trends, public interests, and the general direction of technological advancements. This project utilizes data scraping techniques to gather articles from news sources, followed by data cleaning, analysis, and visualization to extract meaningful information.

III. Methodology

1. Data Collection

The data for this project was sourced from a publicly available API provided by NewsAPI.org. The Python script **get_data.py** was used to fetch news articles related to technology. The API's response, containing article titles, descriptions, and content, was saved in a raw JSON format.

2. Data Cleaning

The raw data was cleaned using the script **clean_data.py**. The cleaning process involved removing HTML tags, handling missing values, and structuring the data into a more analyzable format. Only relevant fields such as titles, descriptions, and content were retained.

3. Data Analysis

Data analysis was performed using the **run_analysis.py** script. Python libraries such as Pandas were utilized to conduct statistical analyses and generate descriptive statistics. The focus was on understanding the distribution and general characteristics of the article data.

4. Data Visualization

The script **visualize_results.py** was employed to create visual representations of the analyzed data. Using Matplotlib, histograms and other plots were generated to visualize aspects like title length distribution, highlighting patterns in the data.

IV. Results

The analysis revealed key characteristics of technology news articles. Descriptive statistics provided insights into the average length of articles, common themes, and the frequency of certain keywords. Visualizations further illustrated these findings, offering a clear representation of data trends.

Gufeng Shen
11/16/2023

V. Discussion

The results indicate a strong focus on specific technology sectors and emerging trends. The data cleaning process was crucial in ensuring the reliability of the analysis. The limitations of the study include the reliance on a single data source and potential biases in the articles chosen by the API.

VI. Conclusion

This project demonstrates the power of Python in extracting and analyzing web data. The insights gained from the technology news articles can be valuable for understanding public interest in technology and identifying emerging trends. Future work could expand the data sources and apply more complex analytical techniques to deepen the understanding of the subject matter.

VII. References

- NewsAPI.org. (Year). API Documentation. https://newsapi.org/docs
- Python Software Foundation. (Year). Python Language Reference, version 3.x. https://docs.python.org/3/reference/index.html
- McKinney, W. (Year). Python for Data Analysis. O'Reilly Media.