edunet foundation

Text Summarization for NEWS Articles

Team Members: Ayushi Vaishnav
Disha Patel
Kashish Patel
Jothsna Sri Kathyayani Chillimuntha

Guide:Praful Vinayak Bhoyar Sir

edunet foundation

OUTLINE

- Abstract
- Problem Statement
- Aims, Objective & Proposed System/Solution
- System Design/Architecture
- System Development Approach (Technology Used)
- Algorithm & Deployment
- Conclusion
- Future Scope
- References
- Video of the Project

edunet foundation

Abstract

This project leverages Natural Language Processing (NLP) techniques to perform comprehensive text analysis and summarization. Using the Natural Language Toolkit (NLTK) for tokenization and stopword removal, TextBlob for sentiment analysis, and Hugging Face's Transformers library for hate speech detection, we analyze and visualize text data from multiple documents.

The process involves reading text files, creating frequency tables to score sentences, and generating concise summaries. Sentiment analysis provides insights into the tone of the summaries, while hate speech detection identifies harmful content. Word clouds are created to visually represent significant terms, showcasing the powerful capabilities of modern NLP tools in processing, analyzing, and summarizing text for applications in content moderation and sentiment analysis.

edunet foundation

Problem Statement

In today's information-rich world, the rapid production and dissemination of news articles create a significant challenge in efficiently extracting meaningful insights. Manual summarization, sentiment analysis, and hate speech detection are time-consuming and prone to error, necessitating automated tools for processing large volumes of text. This project leverages advanced Natural Language Processing (NLP) techniques to develop an automated system that provides concise summaries, sentiment analysis, and hate speech detection, enhancing information management, content moderation, and aiding journalists, editors, and readers in quickly understanding news trends and public opinion.

So, basically ***developing NLP tools for automated news summarization, sentiment analysis, and hate speech detection to aid efficient content processing.***

edunet foundation

Aim and Objective

Aim: To develop an automated system using Natural Language Processing (NLP) techniques for efficient analysis and summarization of news articles.

Objectives:

1. **Automated Summarization:** Develop algorithms for generating concise news article summaries.
2. **Sentiment Analysis:** Implement tools to analyze the sentiment of news articles.
3. **Hate Speech Detection:** Utilize NLP models to detect offensive language in news content.
4. **System Development:** Create a scalable NLP system for efficient news article analysis.
5. **Evaluation:** Evaluate and optimize the system's performance for accuracy and efficiency.

edunet foundation

Proposed Solution

This project aims to develop an NLP system for efficient news article summarization, sentiment analysis, and hate speech detection. Key components include:

1. **Automated Summarization:** Implement algorithms for generating concise news article summaries.
2. **Sentiment Analysis:** Develop tools for analyzing sentiment in news articles, providing insights into public opinion.
3. **Hate Speech Detection:** Utilize NLP models to detect offensive language in news content, enhancing content moderation.
4. **Integration and Scalability:** Create a scalable system architecture to seamlessly integrate these NLP functionalities.

edunet foundation

System Architecture

The system architecture integrates Natural Language Processing (NLP) functionalities for news analysis:

1. **Data Ingestion:** Imports news articles from diverse sources.
2. **Preprocessing:** Cleans and prepares text using tokenization and normalization.
3. **Automated Summarization:** Generates concise article summaries.
4. **Sentiment Analysis:** Determines public opinion and emotional tone.
5. **Hate Speech Detection:** Identifies offensive language in articles.
6. **Integration Layer:** Unifies summarization, sentiment analysis, and hate speech detection.
7. **Scalability:** Handles large data volumes efficiently.
8. **Evaluation and Optimization:** Ensures accuracy and reliability through rigorous testing.
9. **Output:** Provides processed data for further analysis or presentation.

edunet foundation

System Deployment Approach

The project follows an iterative and incremental approach:

1. **Requirements Gathering:** Define stakeholder needs and prioritize system requirements.
2. **Design:** Create a detailed system architecture and integration plan for NLP components.
3. **Implementation:** Develop the system in stages, starting with core functionalities and incrementally adding features.
4. **Testing:** Conduct thorough testing at each stage to ensure functionality, accuracy, and performance.
5. **Deployment:** Deploy the system in controlled environments, iterating based on feedback and testing.
6. **Monitoring and Maintenance:** Implement monitoring for performance and issues, ensuring ongoing system effectiveness and updates.

edunet foundation

Algorithm & Deployment

The project utilizes various algorithms tailored for Natural Language Processing (NLP) tasks:

1. **Automated Summarization:**

Algorithm: TextRank (implemented through NLTK) for generating summaries based on sentence importance.

2. **Sentiment Analysis:**

Algorithm: TextBlob for calculating sentiment polarity and subjectivity scores.

3. **Hate Speech Detection:**

Algorithm: BERT-based model (utilized through the Transformers library) for classifying text into toxic or non-toxic categories.

The project was implemented and executed in a Google Colab environment for development and testing purposes. However, deployment to a production or cloud environment (such as AWS, Azure etc) for public access was not performed.

edunet foundation

Conclusion

"Implementing this NLP system could significantly enhance news analysis efficiency and content moderation capabilities."

This project's development of a robust Natural Language Processing (NLP) system in Google Colab marks a significant step towards enhancing automated news analysis capabilities. By integrating algorithms for summarization, sentiment analysis, and hate speech detection, the project demonstrates potential benefits in improving efficiency and accuracy in content moderation and public opinion monitoring.

edunet foundation

Future Scope

Future deployment of this system could empower news organizations with scalable tools for *effective data handling, enhancing decision-making and promoting secure, insightful content dissemination in digital media landscapes. Enhancements* may integrate advanced deep learning models for sentiment analysis and hate speech detection, explore real-time data streaming and cloud deployment via AWS or Azure for improved scalability, and extend functionalities for multi-language support and context-aware analysis, ensuring robust content moderation capabilities across global media environments.

edunet foundation

Reference

- https://www.oreilly.com/library/view/natural-language-processing/9781787285101/ch27s04.html

edunet foundation

Output:

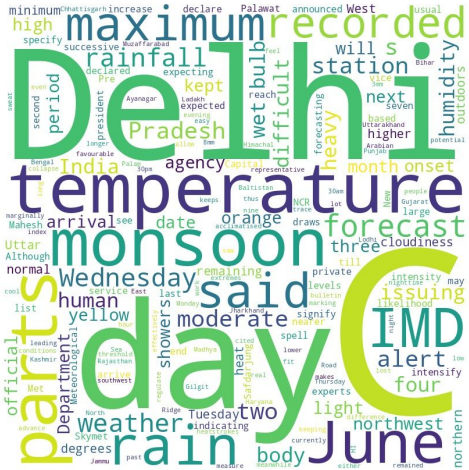
New Delhi: Pre-monsoon showers on Wednesday kept Delhi's maximum temperature below 40°C for the second successive day, but kept humidity levels high, said the India Meteorological Department (IMD). The weather agency also announced the arrival of monsoon over large parts of northwest India over the next three to four days, but the list did not specify a date for Delhi, indicating that monsoon will not arrive on its usual date of June 27. Delhi-NCR may see heavy rainfall spell during this period, based on which the onset of the monsoon could be declared," said Mahesh Palawat, vice president at Skymet, a private weather forecasting service. In its monsoon bulletin on Wednesday evening, IMD said conditions were currently favourable for the southwest monsoon to advance into "the remaining parts of the North Arabian Sea, Gujarat and Madhya Pradesh; some more parts of Rajasthan; remaining parts of Chhattisgarh, West Bengal, Jharkhand and Bihar; most parts of East Uttar Pradesh; some more parts of West Uttar Pradesh; some parts of Uttarakhand, Himachal Pradesh, Jammu-Kashmir-Ladakh-Gilgit-Baltistan-Muzaffarabad, the northern parts of Punjab and northern parts of Haryana during next three to four days."

Delhi recorded a maximum temperature of 39°C on Wednesday, which was two degrees above normal, but the heat index (HI) or "real feel" temperature was 52°C, the same as the past two days. Delhi's humidity remained high, between 51% and 71%, on Wednesday, keeping the wet-bulb temperature — a measure of how difficult or easy it is for the human body to be outdoors — at 29.3°C, which was marginally lower than the 29.6°C recorded on Tuesday.

Sentiment of file1.txt:
Polarity: 0.07972648563557654, Subjectivity: 0.37938410074773715

Hate Speech Detection for file1.txt:
[{'label': 'toxic', 'score': 0.0006439175340346992}]

Word Cloud for file1.txt



Activate Windows

edunet foundation

Thank you!