PYTHON PROGRAMMING PROJECT

EVALUATION -1 REPORT ON

Topic Name: News Aggregator & Summariser

Submitted By: Abhas Jaiswal
Batch: 37 (B.tech CSE)
Semester: II
SAP ID: 500122850
Enrollment Number: R2142231226

Under the guidance of Ms. Gaytri (Assistant Professor) Department of Systems

Department of Systems School of Computer Science

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

Dehradun-248007

Jan-March, 2024

INDEX

Chapter 1: Abstract.. 3

Chapter 2: Introduction..3

2.1: Web Scraping...5

2.2: Natural Language Processing.. 5

Chapter 3: Literature Review...5

Chapter 4: Inference from Literature Review..6

Chapter 5: Objectives..8

Chapter 6: Analysis..8

Chapter 7: Proposed System.. 10

Chapter 8: Conclusion...11

References...12

Chapter 1: Abstract

In today's digital age, the abundance of online news sources presents users with the challenge of navigating through a vast volume of information. News aggregators have emerged as essential tools for simplifying access to diverse news content, yet users often encounter content overload, hindering their ability to efficiently consume news. This paper builds a news aggregator from scratch using web-scraping and explores the integration of natural language processing (NLP) techniques to address the challenge of content overloading. By leveraging web scraping for data collection and NLP for text summarization, our proposed system aims to streamline the news aggregation process, providing users with concise, relevant summaries of news articles. Through a comprehensive review of existing literature, this research identifies key challenges in news aggregation and proposes a novel approach to enhance the user experience, ultimately contributing to the advancement of news consumption in the digital era.

Chapter 2: Introduction

The landscape of news consumption has undergone a significant shift in recent years, propelled by the advent of digital technologies and the ubiquity of the internet. Traditional modes of news dissemination, such as newspapers and television broadcasts, have given way to the dynamic and interconnected realm of digital news platforms. According to a recent survey by the Pew Research Center [1], a staggering 93% of people get at least some news online, marking a significant departure from conventional print and broadcast media. As per another survey by Reuters Institute for the Study of Journalism [2], digital news consumption has surpassed traditional television news across several key demographics.

The allure of digital news lies in its immediacy, interactivity, and accessibility, allowing users to engage with a diverse array of perspectives and sources at their fingertips. However, with the abundance of news websites [3 - 6] and the vast amount of information available, it can be overwhelming and time-consuming to visit multiple sources to gather news. This is where a news content aggregator comes into play. A news aggregator [7 - 11] simplifies this process by allowing users to select the websites that users want to follow, then the aggregator collects articles from these sources, enabling users to access information from multiple websites with a single click. This not only saves time but also provides users with a competitive advantage by granting them valuable knowledge that others may lack.

However, alongside their benefits, there are growing concerns regarding the quality, reliability, and transparency of digital news content and the platforms that deliver it. As users navigate through a vast sea of information, they encounter a myriad of challenges,

1. from distinguishing between credible sources and misinformation to,
2. grappling with content overload.

These issues underscore the need for a critical examination of the current state of digital news and the role of news aggregators in shaping our information ecosystem.

To tackle the problem of ensuring authenticity and credibility, news aggregators use trustworthy sources like major news outlets, blogs, and online publications. They cover various topics like politics, sports, and technology, giving users a wide range of news to choose from. This helps people find reliable news easily and saves them from sorting through less credible sources online.

This research aims to tackle the problem of content overload faced by news aggregators. Many aggregators simply gather articles from various sources without organizing them properly, leading to users seeing multiple duplicates of the same news. However, recent advancements in natural language processing (NLP) [12] technology offer a solution. NLP can automatically analyze large volumes of articles, grouping together those covering the same topic and providing a concise summary. This helps users get a clearer picture of the news without being overwhelmed by duplicates. As part of the research project, we build a news-aggregator from scratch and focus mainly on tackling the issue of content overloading.

The final news-aggregation system consists of two major steps:

1. Web Scraping
2. Natural Language Processing

2.1: Web Scraping

The aggregation process generally involves utilizing web crawling [13] and parsing techniques to extract news articles from predefined sources. Web scraping is a crucial technique used in news aggregation to gather articles from various online sources. It involves automatically extracting information from web pages, allowing aggregators to access a wide range of news content. Through web scraping, aggregators can retrieve articles from news websites, blogs, and other trusted platforms, ensuring a comprehensive coverage of news topics. By programmatically navigating through web pages and extracting relevant data such as article headlines, text, and publication dates, web scraping enables aggregators to compile a diverse array of news articles for users to access in one centralized location. This process streamlines the collection of news content, facilitating efficient and timely updates for users seeking the latest information across different topics.

2.2: Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It encompasses a range of tasks including sentiment analysis, language translation, and text summarization [12]. Recent advancements in NLP have revolutionized various industries, including news aggregation. NLP techniques allow aggregators to automatically analyze and categorize news articles, extract key information, and even generate summaries, making it easier for users to access relevant content [13]. By leveraging NLP, news aggregators can enhance the accuracy and efficiency of their platforms, providing users with personalized and streamlined news consumption experiences.

In short, news aggregators are handy tools for staying updated in today's digital age. They simplify the process of finding trustworthy news and provide a one-stop platform for accessing a variety of news articles from reputable sources. The rest of the paper is organized as follows: Chapter 3 provided a review of the existing work, Chapter 4 provides an insight into the existing work and issues with the existing work, Chapter 5 defines the objectives of this research work, Chapter 6 talks about the performance metrics associated with the existing solutions, Chapter 7 details the proposed system and the paper concludes with a conclusion in Chapter 8.

Chapter 3: Literature Review

In the rapidly evolving digital landscape, news aggregation platforms [3 - 6] have emerged as indispensable tools for individuals seeking convenient access to current events from a multitude of online sources. These platforms, often powered by web scraping techniques, enable users to gather news articles from diverse sources and curate them into a centralized hub for easy access and consumption.

Web scraping plays a pivotal role in the aggregation process, facilitating the automated collection of news articles from various online sources. Bhujbal et al. [14] demonstrated the effectiveness of web scraping in their study, where they utilized scraping techniques to gather news articles and categorize them based on topic or relevance. By programmatically accessing and parsing web pages, web scraping allows aggregators to compile a comprehensive selection of news articles covering a wide range of topics.

However, the abundance of news articles collected through web scraping can overwhelm users, necessitating the integration of techniques for effective information summarization. Natural Language Processing (NLP) techniques offer a promising solution to this challenge by enabling the automatic summarization of news articles. La Quatra et al. [15] developed an NLP-based text summarization model that extracts salient information from texts and generates concise summaries. Their research highlighted the importance of considering factors such as sentence relevance, coherence, and informativeness in the summarization process.

Moreover, NLP can be leveraged for multi-document summarization, wherein information from multiple articles on the same topic is synthesized to produce a comprehensive summary. This approach enables users to gain a holistic understanding of complex news topics by identifying common themes and extracting key points from disparate sources.

Despite the potential benefits of integrating NLP for text summarization in news aggregation platforms, several challenges must be addressed. Scalability, accuracy, and computational efficiency are key considerations in developing robust summarization algorithms that can handle large volumes of data in real-time. Furthermore, ensuring the accuracy and reliability of summaries requires careful validation and evaluation against human-generated summaries.

In conclusion, the combination of web scraping and NLP holds great promise for enhancing news aggregation platforms. By automatically collecting and summarizing news articles from diverse sources, these platforms empower users to efficiently navigate and digest vast amounts of news content, thereby facilitating informed decision-making and fostering broader societal engagement.

Chapter 4: Inference from Literature Review

1. **News Aggregation Platforms as Integral Tools:** The literature review underscores the significance of news aggregation platforms as indispensable tools in the digital landscape. These platforms, often utilizing web scraping techniques, enable users to conveniently access current events from a multitude of online sources [3 - 6].

2. **Role of Web Scraping in Aggregation:** Web scraping emerges as a pivotal technique in the aggregation process, facilitating the automated collection of news articles from various online sources. Bhujbal et al. [14] demonstrate the effectiveness of web scraping in gathering and categorizing news articles based on topic or relevance.

3. **Challenges of Content Overload:** While web scraping enables comprehensive news coverage, the abundance of collected articles can overwhelm users. This necessitates the integration of techniques for effective information summarization, as highlighted by the study.

4. **Promise of NLP in Summarization:** Natural Language Processing (NLP) techniques offer a promising solution to the challenge of content overload by enabling automatic summarization of news articles. La Quatra et al. [15] demonstrate the development of an NLP-based text summarization model, emphasizing the importance of factors such as sentence relevance, coherence, and informativeness in the summarization process.

5. **Benefits of Multi-Document Summarization:** NLP can also be leveraged for multi-document summarization, synthesizing information from multiple articles on the same topic to produce comprehensive summaries. This approach enhances users' understanding of complex news topics by identifying common themes and extracting key points from disparate sources.

6. **Challenges in Algorithm Development:** Despite the potential benefits, challenges such as scalability, accuracy, and computational efficiency need to be addressed in developing robust summarization algorithms that can handle large volumes of data in real-time. Additionally, ensuring the accuracy and reliability of summaries requires careful validation and evaluation against human-generated summaries.

7. **Promise for Enhanced User Experience:** In conclusion, the combination of web scraping and NLP holds great promise for enhancing news aggregation platforms. By automating the collection and summarization of news articles from diverse sources, these platforms empower users to efficiently navigate and digest vast amounts of news content, thereby facilitating informed decision-making and fostering broader societal engagement.

Chapter 5: Objectives

1. Understanding web scraping and its application to obtain news data from various news portals.
2. To identify key elements to scrape from different news portals based on their structure and content. In addition to headlines, summaries, and authors, publication dates are also included in these elements.
3. Automating news article extraction from multiple news portals using web scraping techniques and tools.
4. To explore data processing and integration techniques to organize and present scraped news data in a coherent and user-friendly manner on an aggregated news platform.
5. Use topic-modeling NLP models, to tag or associate a category with all the scraped articles.

6. Use text-summarization NLP models to summarize articles/news collected from multiple sources about the same or similar incident.

Chapter 6: Analysis

Existing news aggregation platforms play a vital role in providing users with convenient access to a wide range of news sources and topics. These platforms utilize various techniques such as web scraping, natural language processing (NLP), and user customization features to enhance the news browsing experience.

Note: As there is no numeric metric to quantify the performance, a detailed qualitative analysis of the existing news-aggregation platforms has been conducted on six different quality metrics:

1. **Key Features:**
    - **a.** Web Scraping: Most platforms employ web scraping techniques to gather news articles from multiple online sources. This ensures comprehensive coverage and up-to-date information.
    - **b.** Personalization: Many platforms offer customization options, allowing users to select their preferred news sources and topics of interest. This feature enhances user engagement and satisfaction.
    - **c.** Category Filtering: Platforms often categorize news articles into topics such as politics, technology, sports, etc. Users can filter articles based on these categories to find content relevant to their interests.

2. **Popular Platforms:**
    - **a.** Google News: Google News aggregates news articles from various sources and provides personalized recommendations based on user interests and browsing history.
    - **b.** Flipboard: Flipboard offers a visually appealing interface where users can create personalized "magazines" by selecting their favorite topics and sources.
    - **c.** Feedly: Feedly allows users to follow their favorite websites and blogs, organizing content into customizable feeds for easy browsing.
    - **d.** Apple News: Apple News curates news articles from trusted sources and offers personalized recommendations based on user preferences.

3. **Strengths:**
    - **a.** Comprehensive Coverage: Existing platforms offer a wide range of news sources and topics, ensuring users have access to diverse perspectives.

   - **b.** User-Friendly Interface: Many platforms prioritize user experience with intuitive interfaces and customizable features.
   - **c.** Personalization: The ability to customize news feeds based on user preferences enhances user engagement and satisfaction.

4. **Weaknesses:**
   - **a.** Quality Control: With the abundance of news sources, ensuring the quality and reliability of information can be challenging.
   - **b.** Content Overload: Some users may feel overwhelmed by the sheer volume of news articles, leading to difficulty in finding relevant content.
   - **c.** Privacy Concerns: Platforms that personalize recommendations based on user data may raise privacy concerns among users.

5. **Opportunities:**
   - **a.** Advanced NLP Integration: Further integration of NLP algorithms for content summarization and sentiment analysis can enhance the value proposition of news aggregation platforms.
   - **b.** Community Engagement: Incorporating social features such as comments, likes, and shares can foster community engagement and user interaction.
   - **c.** Cross-Platform Integration: Seamless integration with other digital platforms such as social media or messaging apps can broaden the reach and accessibility of news content.
6. **Threats**:
   - **a.** Competition: The news aggregation market is highly competitive, with new entrants constantly emerging. Established platforms must innovate to maintain their competitive edge.
   - **b.** Fake News: The proliferation of fake news poses a significant threat to the credibility and trustworthiness of news aggregation platforms. Platforms must implement robust measures to combat misinformation and ensure the integrity of their content.

Existing news aggregation platforms offer valuable services to users by providing convenient access to news content from diverse sources. By leveraging technology and user customization features, these platforms enhance the news browsing experience. However, they also face challenges such as maintaining quality control and combating content overload. Moving forward, opportunities for innovation and improvement lie in advanced NLP integration, community engagement, and cross-platform integration. Platforms must remain vigilant against threats such as competition and fake news to uphold their credibility and relevance in the digital age.

Chapter 7: Proposed System

The proposed system is a Python-based news aggregator that aims to streamline the process of accessing and consuming news articles from various online sources. By leveraging web scraping techniques, the system fetches news articles from predefined websites, allowing users to customize their news consumption experience. Additionally, the integration of the Google GenAI API enables the generation of summarized versions of news articles, enhancing comprehension and efficient browsing. Figure 1. Represents the key components of the proposed system.

Key Components of the Proposed System:

Website Scraping: The system utilizes libraries such as Selenium, Beautiful Soup, and Requests to scrape news articles from predefined websites. This process involves making HTTP requests and employing browser automation to extract data from web pages.

Source Selection: Users have the flexibility to select their preferred news sources from a list of supported websites. This feature enables users to curate their news feed according to their interests and preferences.

Category Filtering: The system allows users to filter news articles based on predefined categories such as politics, technology, sports, etc. This feature enhances user experience by enabling targeted content consumption.

Article Summarization: Leveraging the Google GenAI API, the system generates summarized versions of news articles. This advanced natural language processing capability provides users with concise summaries, facilitating quick comprehension and efficient browsing of multiple news stories.

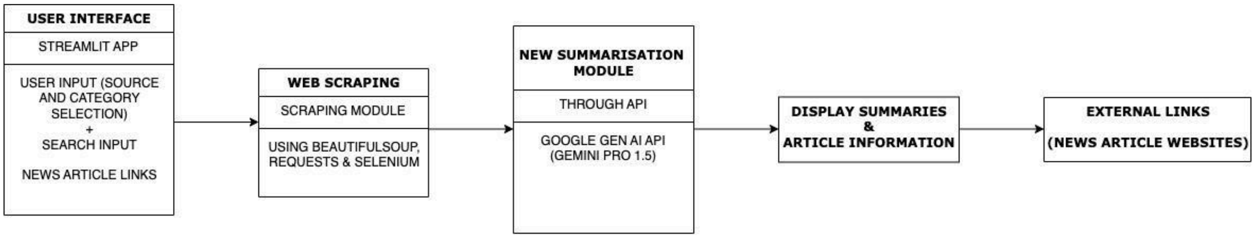

Figure 1: The key components of the proposed python-based news-aggregator with news summarisation powered by NLP.

Chapter 8: Conclusion

In conclusion, this paper has delved into the evolving landscape of news consumption in the digital age, highlighting the challenges posed by content overload and the role of news aggregators in addressing these challenges. The integration of natural language processing (NLP) techniques offers a promising solution to streamline the news aggregation process and enhance the user experience. By leveraging web scraping for data collection and NLP for text summarization, our proposed system aims to provide users with concise, relevant summaries of news articles, thereby alleviating the burden of content overload.

Through a comprehensive review of existing literature and the development of a novel approach to news aggregation, this research contributes to the advancement of news consumption in the digital era. By automating the process of article summarization, our system empowers users to make informed decisions without being overwhelmed by excessive information. Moving forward, further research and development in NLP technologies will continue to shape the future of news aggregation, ultimately facilitating a more efficient and engaging news consumption experience for users worldwide.

References

[1] Pew Research Center, "News Use Across Social Media Platforms 2021," Pew Research Center, 2021. [Online]. Available: https://www.pewresearch.org/journalism/2021/07/27/news-use-across-social-media-platforms-2021/.

[2] Reuters Institute for the Study of Journalism, "Digital News Report 2023," Reuters Institute for the Study of Journalism, 2023. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

[3] “The Times of India” : https://timesofindia.indiatimes.com/

[4] “Hindustan Times”: https://www.hindustantimes.com/

[5] “The New York Times”: https://www.nytimes.com/section/todayspaper

[6] “The Guardian”: https://www.theguardian.com/international

[7] “Google News”: https://news.google.com/home?hl=en-IN&gl=IN&ceid=IN:en

[8] “Flipboard”: https://flipboard.com/

[9] “Feedly”: https://feedly.com/

[10] “SmartNews”: https://www.smartnews.com/

[11] Web-crawlers: https://en.wikipedia.org/wiki/Web_crawler

[12] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.

[13] Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. Journal of Artificial Intelligence Research, 57, 345-420.

[14] Bhujbal, Mayur & Bibawanekar, Ms & Deshmukh, Pratibha. (2023). News Aggregation using Web Scraping News Portals. International Journal of Advanced Research in Science, Communication and Technology. Volume 3. 2581-9429. 10.48175/IJARSCT-12138.

[15] La Quatra, M. (2022). Deep Learning for Natural Language Understanding and Summarization (Doctoral dissertation, Politecnico di Torino).