Names: Kanchan Ashok Naik **ID:** 016729128
Mrunali Katta 017516785
Prasad Pramod Shimpatwar 016722251
Yashasvi Kanchugantla 017521699

Wikinews Insights: Real-time Correlation and Summarization of Wikipedia Pageviews and News Headlines

Problem Statement:

The aim of this project is to implement a "WikiNews Insights" system that combines real-time Wikipedia pageview data with current news headlines to provide insights into public interest and information seeking behavior. It monitors statistics across multiple language editions, aggregates and analyzes news headlines, and provides concise summaries of correlated articles and news stories. The system ensures user's privacy, updates in real-time, and presents findings in an accessible dashboard.

Abstract:

The "WikiNews Insights" tool is designed to create a real-time analytical link between Wikipedia pageviews and current trending news headlines, offering fresh insights into public interest and their behaviors. Its main objective is to develop a scalable big data pipeline that processes Wikipedia data hourly to identify trending topics, which are then connected to the latest news articles sourced from major news APIs. This system aims to reveal the correlation between public engagement and current events, presenting the results through easy-to-use, real-time visualizations and interactive dashboards, etc. By utilizing advanced data processing technologies and summarization algorithms, the project not only enhances the understanding of digital content consumption but also supports various applications, from journalistic research to academic studies. This approach marks a new era in media analytics, where real-time data improves the responsiveness and informed nature of media distribution.

Motivation:

Wikipedia is the most used web encyclopedia there is. Wikipedia has projects in 337 languages and are being actively maintained. There are 61M pages and about 238M pageviews[2] are observed just on en.wikipedia.org. It is the source of all primary information one needs to start with. In this project, we want to understand the relationship between public interest(as reflected in Wikipedia pageviews) and current events in the news.

This project uniquely combines Wikipedia pageview analysis with news aggregation, creating a novel approach to understanding public interest. Solving this use case is a good opportunity to apply several Big Data components learned in this course. Integrating different data sources helps make it useful for social researchers, journalists, and media persons.

Literature Survey:

The base idea of this project is inspired by the research article [5] which discusses the potential of Wikipedia pageviews as a measure of investor interest and their relationship with the performance of the Nasdaq index. By utilizing big data methodologies, this article develops two econometric models to assess the explanatory and predictive capabilities of Wikipedia visits concerning stock returns. While Wikipedia traffic is found to significantly mirror Nasdaq trends, it does not possess the ability to predict future market movements. The study advocates for additional exploration into the profitable application

of this indicator within financial markets, highlighting the integration of big data and algorithmic trading as promising directions for future research.

Methodology:

Scope

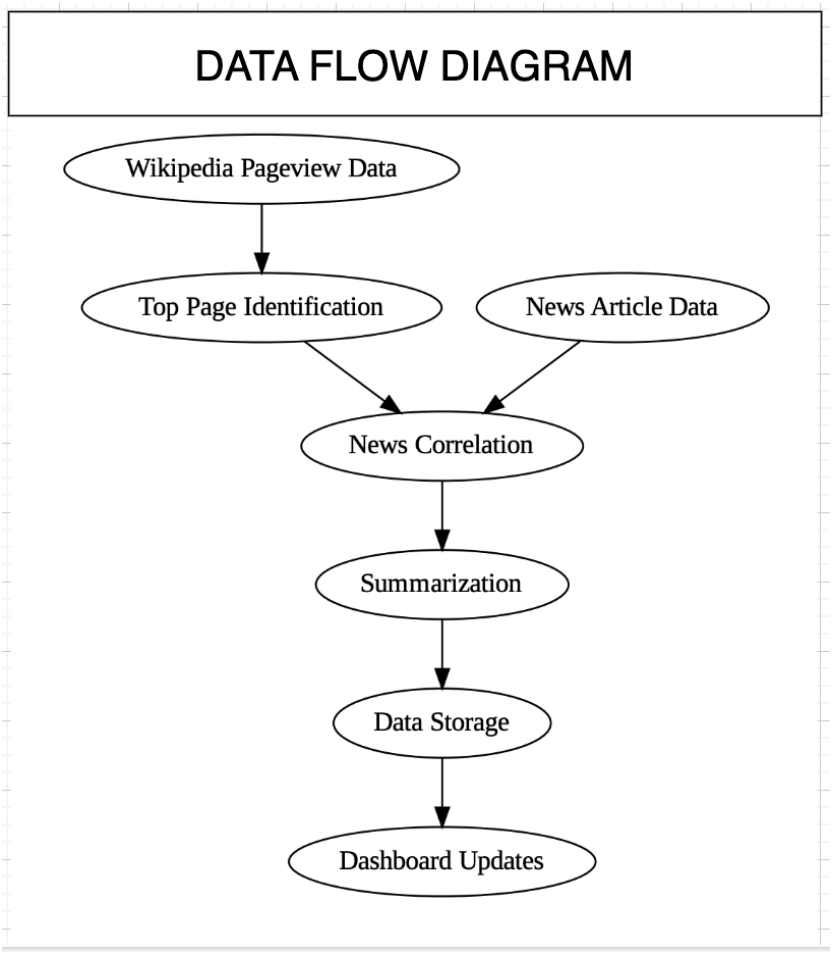
1. Create a data pipeline to collect and process Wikipedia pageview statistics and news headlines.
2. Develop **algorithms to identify and quantify correlations between spikes** in Wikipedia pageviews and trending news topics.
3. Create **concise, informative summaries of correlated Wikipedia article**s **and news stories** to provide key information related to trending topics.
4. Design and develop an intuitive, interactive dashboard that visualizes pageview trends.
5. Validate the accuracy and reliability of the system's correlations and summaries through rigorous testing and comparison with known events.
6. Contribute to the broader understanding of digital information consumption patterns and their relationship to real-world events.

Data Collection

- Wikipedia - Wikipedia pageview data using the Wikimedia EventStreams API (https://stream.wikimedia.org/v2/stream/pageviews)
- We will be using programmatic access using Python to Wiki statistics using various APIs provided by Wikipedia and statistics calculated from data for the past one hour will be stored in the system.
- News Headlines and Top News data from News APIs (https://newsapi.org/v2/everything?q=Erik-Menendez&X-Api-Key=<APIKEY>
- News Headlines and Top News from Currents (https://api.currentsapi.services/v1/latest-news?language=en&apiKey=<APIKEY>
- Extract and process raw JSON.
- Web Crawling to fetch content of original news articles from news websites for news summarization.

Data Pre-Processing

1. Data Extraction
    a. Extract relevant fields from News articles and Wiki logs.
    b. Extract main content from crawled web pages without advertisements and page elements.
2. Data Cleaning
    a. Remove invalid entries like internal logs or where the title or data source is not mentioned.
    b. Remove duplicate data from logs and news articles.
    c. Aggregate Wikipedia data by page views and categories.
3. Data Extraction
    a. Extract relevant fields from News articles and Wiki logs.
    b. Extract main content from crawled web pages without advertisements and page elements.
4. Text Cleaning
    a. Remove HTML tags or other special characters from News articles.
    b. Remove other language text from the article.
5. Standardize Date and Time to UTC or common time zone.



Data Processing

1. Tokenize the text into words or subwords.
2. Remove stop words
3. Apply stemming or lemmatization to reduce words to their base forms
4. Perform named entity recognition to identify people, places, organizations, etc.
5. Extract Topic, Categories, and Sentiment Score
6. Correlation coefficient between wiki page view spike and current news trends.
7. Structure the data for efficient storage and retrieval.
8. Prepare data in a specific format for Visualization and Analytics

News Wiki Correlation

1. Implement algorithms to identify spikes in Wikipedia pageviews
2. Calculate correlation coefficients between pageview spikes and relevant news topics
3. Fetch Top news articles from News websites and display most relevant news

Content Summarization

1. Generate summary of news articles and wiki article to highlight current event.
2. We will be using extractive summarization techniques like using TextRank or LexRank algorithms
3. Along with extractive summarization, we will explore Abstractive summarization using pre-trained transformer models.

Data storage and Visualization

1. We will be storing summarized data and wiki spike related information in distributed database such as MongoDB or Apache Cassandra.
2. Visualization and summaries will be displayed using an application endpoint built using Python or Django.

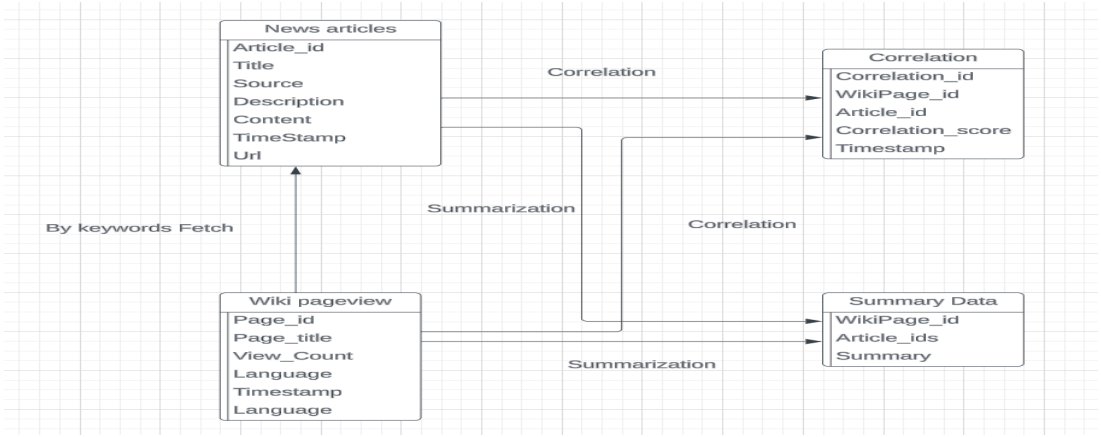
Tools and Techniques

1. Apache Kafka will be used to stream wiki stream for real-time streaming
2. Apache spark to processes wiki page views and identify top pages
3. Map-Reduce will be used to aggregate and categorize news articles and pageviews
4. Further, we will employ Map-Reduce to filter dataset based on keywords
5. We will be using Bloom Filter to check if data or pageview is already processed or not to avoid duplicate processing.
6. Once data is deduplicated we can use the Reservoir sampling algorithm to sample pageviews from the stream of data to maintain a representative sample of data, even as the total number of page view grows indefinitely. This will allow us to perform analysis without storing actual data.

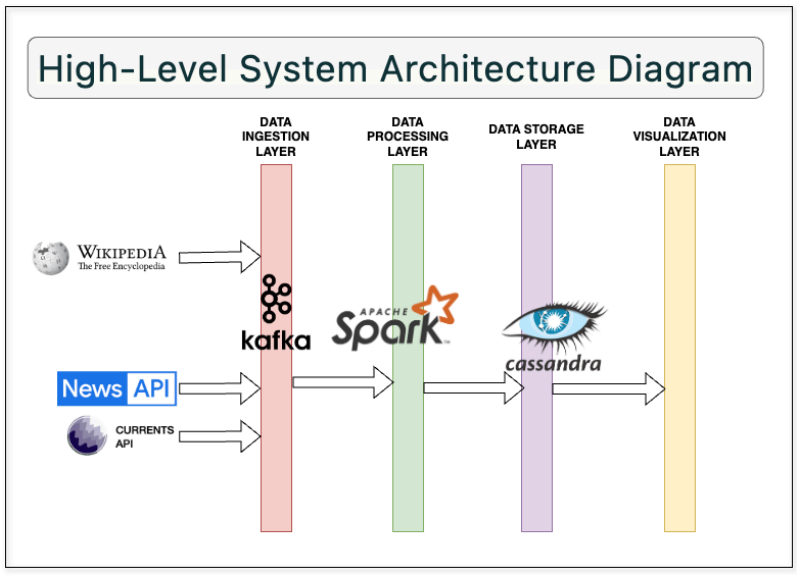
Evaluation Methods

1. **Human Evaluation** - For summarization of news topics we will be employing human evaluation. i.e we will be checking ourselves if a generated summary is of good quality or not.
2. **Precision and Recall -** We will be checking how many correlated articles are correctly correlated with wiki page views.
3. **Temporal Alignment Accuracy -** We will be checking page view spike and Time of News article publication aligns or not.
4. **Performance and Latency -** We will be using quality testing metrics like throughput and load capacity of the system using tools such as Jmeter.

Data Model



Architecture Diagram



Deliverables:

1. A comprehensive report summarizing the project's development, key technologies, and methodologies.
2. We plan to implement dynamic visualizations to depict real-time data trends through various graphs and plots
3. The web application, to display visualizations and summarized news articles.
4. Final presentation deck for project.
5. A research paper describing methodology and findings

Team Members and Roles:

1. Implementation to ingest stream data from wiki	Prasad Pramod Shimpatwar
2. Bloom filter and reservoir filter for sampling	Mrunali Katta
3. Data cleaning using map-reduce	Kanchan Ashok Naik
4. Correlation evaluator	Yashasvi Kanchugantla
5. Summarizer	Prasad Pramod Shimpatwar
6. Data Aggregation and processing (Tagging, categorizing other required things)	Mrunali Katta
7. News and currents api call and filtering relevant results	Kanchan Ashok Naik
8. Web crawling to fetch and extract data from original news articles	Yashasvi Kanchugantla
9. UI desktop in Django for displaying output	Prasad Pramod Shimpatwar

Relevance to the course

1. The project involves processing large-scale Wikipedia pageviews and news headlines, connecting to the course's focus on Big Data tools like Hadoop and Spark.
2. It requires stream processing techniques, such as Spark Streaming and Kafka, to handle continuous real-time data streams.
3. Algorithms like Bloom Filters and Reservoir Sampling for managing data efficiently relate to course topics on stream processing algorithms like Flajolet-Martin and DGIM.
4. Algorithms for mining patterns, such as PageRank and Locality Sensitive Hashing will helps us in identifying correlations between Wikipedia pageviews and trending news mirrors .
5. Providing real-time insights aligns with the course's emphasis on the importance of **stream processing frameworks** and **algorithms for real-time data processing** (like Spark and Kafka) for timely and efficient decision-making.

Technical Difficulty:

Managing High-Throughput, Real-Time Data Pipelines: Continuous ingestion of large volume of data from Wikipedia pageviews and news source and managing and keeping track of this data we might need powerful system and data pipelines

Seamlessly Integrating Multi-Source Data Feeds: Injecting data from diverse APIs—such as Wikipedia, news platforms requires handling different type of data formats synchronization between these sources while managing rate limits complex while processing.

Building Advanced NLP-Based Text Summarization: Combining large amounts of unstructured text from news articles and Wikipedia pages accurately requires advanced Natural Language Processing (NLP) techniques to generate sensible and logical summaries that retain essential user information.

Ensuring Minimal Latency and Maximizing Performance: Real-time data processing requires minimal latency in data ingestion,conversion, and analysis. Delays in this workflow can reduce the value of data and user experience.

Personalized Insight Generation and User-Specific Recommendations: Users may have different preferences or interest areas, and the system must be capable of delivering personalized, relevant insights based on their requirement and area of interest

Novelty:

The "WikiNews Insights" system statistically **predicts viral topics** by analyzing different patterns in Wikipedia traffic before they are reported by mainstream news sources providing important warnings for market trends and social movements. It applies **Bloom Filters** for real-time deduplication and **Reservoir Sampling** to optimize large data streams, making it fast and scalable for processing. The integration of these techniques allows the system to handle huge volume of data while maintaining accuracy. By combining **real-time Wikipedia pageviews** with different news sources improves the understanding of how public interest correlates with current news coverage. This approach offers deeper insights into public engagement with real-world events.

Impact:

With the "WikiNews Insights" analysis, we can help various businesses like the media professionals aka journalists, media houses to improve public understanding of the relationship between the current news events and the information-seeking behavior on Wikipedia. It will provide real-time data on the correlation between Wikipedia pageviews and news headlines, allowing users to see which news events are currently influencing the public interest and search activities, thus enhancing the public engagement with the real time events.

Heilmeier Catechism:

1) **What are you trying to do?**

   Develop a system that correlates real-time Wikipedia pageview data with current news headlines to provide timely insights into public interest and information-seeking behavior. The system will identify correlations between spikes in pageviews and trending news topics, generating brief summaries and visualizations that illustrate these relationships.

2) **How is it done today, and what are the limits of current practice?**

   Most studies on public interest and media trends use separate data sources like social media or news articles. Many tools check social media activity or how news articles perform, but they often ignore Wikipedia pageviews. This means they can miss important information.

3) **What's new in your approach and why do you think it will be successful?**

   The "WikiNews Insights" approach combines real-time data from Wikipedia pageviews and trusted news sources. It predicts viral topics using Wikipedia data, processes large data efficiently with

techniques like Bloom Filters, and analyzes trends in regions. This method aims to understand global and local public interests effectively.

4) **Who cares?**

Researchers can use the data to study trends in public interest and behavior. **Journalists** can improve their reporting by knowing what topics are popular with the public. **Businesses** can analyze market trends to make better decisions. **Public policy makers** can understand public feelings, helping them create policies that meet the needs of the people.

5) **If you're successful, what difference will it make?**

The initiative aims to improve how the media interacts with the public by giving journalists and media organizations real-time insights to make their content more precise and relevant. Last but not least we plan to support research and analysis by providing a detailed dataset as per user requirement for academic studies and public policy discussions.

6) **What are the risks and the payoffs?**

Even if user data is anonymized, people might still be worried about how their information is managed. Relying on automated systems for data processing which can sometimes lead to mistakes or misunderstandings.

7) **How much will it cost?**

We are using free tier resources

8) **How long will it take?**

Planning and Design Stage 1-2 weeks, Development Stage 1-2 months, Testing Stage 2-3 weeks

9) **What are the midterm and final "exams" to check for success?**

The **Midterm Evaluation** will focus primarily on data gathering, data cleaning, and implementing data pipeline's functionality.
The **Final Evaluation** will provide full processing capabilities, accuracy, user acceptance, and adherence to security and privacy standards, with end-user feedback determining if the system achieves its goals and offers valuable insights.

References:

1. https://github.com/public-apis/public-apis?tab=readme-ov-file#news
2. https://arxiv.org/pdf/1509.02218
3. https://api.currentsapi.services/v1/latest-news?language=en&apiKey=OjH7KRQVmCTjUkiU5n2bsB-W8iTQcA7i5AHjjWzhDHts0LJ_
4. https://newsapi.org/v2/everything?q=Erik-Menendez&X-Api-Key=f049190b6b744976b46acc21a25972f9
5. https://onlinelibrary.wiley.com/doi/full/10.1002/isaf.1508