University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

News Article Data Pipeline Technical Report

CONTENTS

News Article Data Pipeline

Executive Summary

This project addresses the challenges posed by the 24-hour news cycle, particularly the difficulty in monitoring trends and identifying relevance. Our pipeline automates the process of news data collection, storage, analysis, and presentation. Key highlights include:

- **Data Ingestion**: Web scraping with Python (BeautifulSoup, Requests).
- **Data Storage**: AWS S3 for scalable and secure storage.
- **Data Analysis**: Keyword scraping to uncover key themes and trends.
- **Deployment**: Serverless architecture using AWS Lambda and monitoring with AWS CloudWatch.

The project's implementation showcases an efficient and scalable architecture capable of processing and analyzing real-time news data.

Team Members:
Thomas Powell
Simran Sattar
Pavan Kumar Sunkara

News Article Data Pipeline

Highlights of Project

Staying updated with relevant news amid the flood of information. An automated pipeline leveraging web scraping, cloud storage and topic modeling. Combination of serverless computing and machine learning. Identifying news trends for journalists, researchers and general users.

Submitted on: 12/08/24

Abstract

In an era dominated by a continuous influx of information, staying informed about current events has become increasingly challenging. This project addresses the problem by developing an automated News Article Data Pipeline to extract and analyze trending news topics efficiently. The system employs Python-based web scraping to retrieve headlines from the Google News homepage, followed by an AWS S3 pipeline for secure storage of the extracted data.

Leveraging keyword tracking and analysis, the project identifies prevalent themes within the headlines, offering a systematic approach to understanding news trends. The architecture is designed with serverless computing using AWS Lambda, ensuring scalable and automated execution. AWS CloudWatch facilitates real-time monitoring and error handling, ensuring reliability and efficiency.

This end-to-end solution provides insights into daily news trends, paving the way for future integration with analytical dashboards and applications. The project not only showcases a scalable and cost-effective architecture but also highlights the potential of combining data science methodologies with cloud-based automation for impactful information retrieval and analysis.

Introductory Section

The rapid pace of information dissemination through the 24-hour news cycle has made it increasingly difficult for individuals to remain informed about relevant and significant current events. This challenge is exacerbated by the overwhelming volume of news published across diverse platforms. To address this issue, our project, the "News Article Data Pipeline," was developed to automate the process of extracting, analyzing, and presenting trending news topics efficiently. Leveraging modern data science techniques, including web scraping, cloud-based storage, and machine learning, this solution provides a systematic approach to understanding and monitoring news trends in real-time.

Review of available research

Automated topic extraction and trend analysis are areas of growing interest in data science. Keyword analysis has been extensively utilized in academic and industry research for text analysis and uncovering latent topics within large datasets. Recent advancements in cloud computing and serverless architectures, such as AWS Lambda, have further enabled scalable and cost-effective solutions for processing large amounts of text data. Research has demonstrated the efficiency of these methods for applications like sentiment analysis, market trend monitoring, and news aggregation. However, there remains a gap in their application to real-time news trend analysis that ensures data reliability and interpretability. Our project bridges this gap by integrating these technologies into a cohesive pipeline to analyze daily news trends effectively.

Methodology:

The project follows the CRISP-DM methodology:

Business Understanding:

Tracking news trends and identifying relevant topics efficiently to support better decision-making.

Data Understanding:

Data is scraped from Google News headlines. Example headlines include:

- *"Trump attends Notre Dame Cathedral reopening in Paris"*
- *"Syrian rebels begin to encircle Damascus"*

Data Preparation:

- Tools: Python libraries for web scraping, AWS services for data handling.
- Pipeline: Extracted data stored in AWS S3 for secure and scalable access.

Modeling:

- Method: Keyword Extraction and Analysis
- Keywords: Identified from text data, including "Notre Dame," "Damascus," "rebels," and "Assad"(keyword_analysis_202412…).

Evaluation:

Results evaluated based on:

- Coherence of topics extracted.
- Timeliness and relevance of stored data.

Deployment:

Automated execution using AWS Lambda and monitoring with AWS CloudWatch.

Results Section

Findings

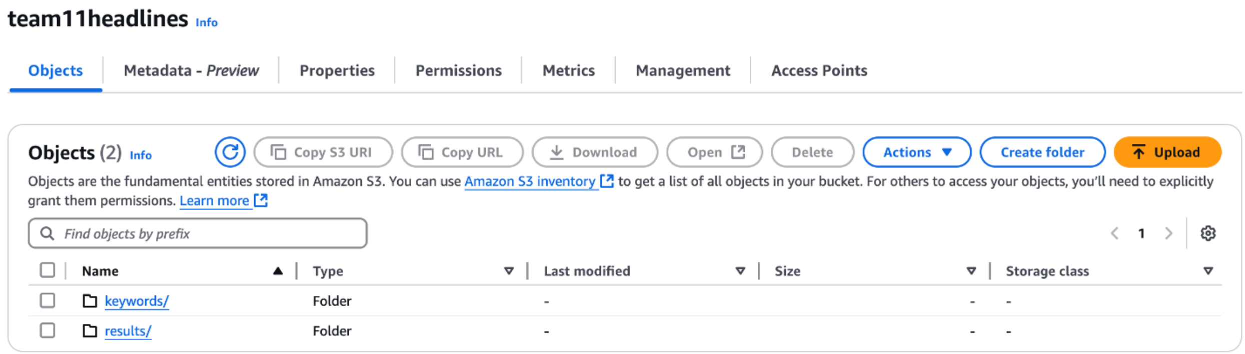
Key topics identified include:

1. **Notre Dame Reopening**:
    - Headlines mention the reopening event attended by world leaders.
    - Keywords: "Notre Dame," "Paris," "cathedral," "reopening".
2. **Syrian Conflict**:

- ○ Topics highlight escalating tensions with Syrian rebels approaching Damascus.
- ○ Keywords: "Damascus," "rebels," "Assad".

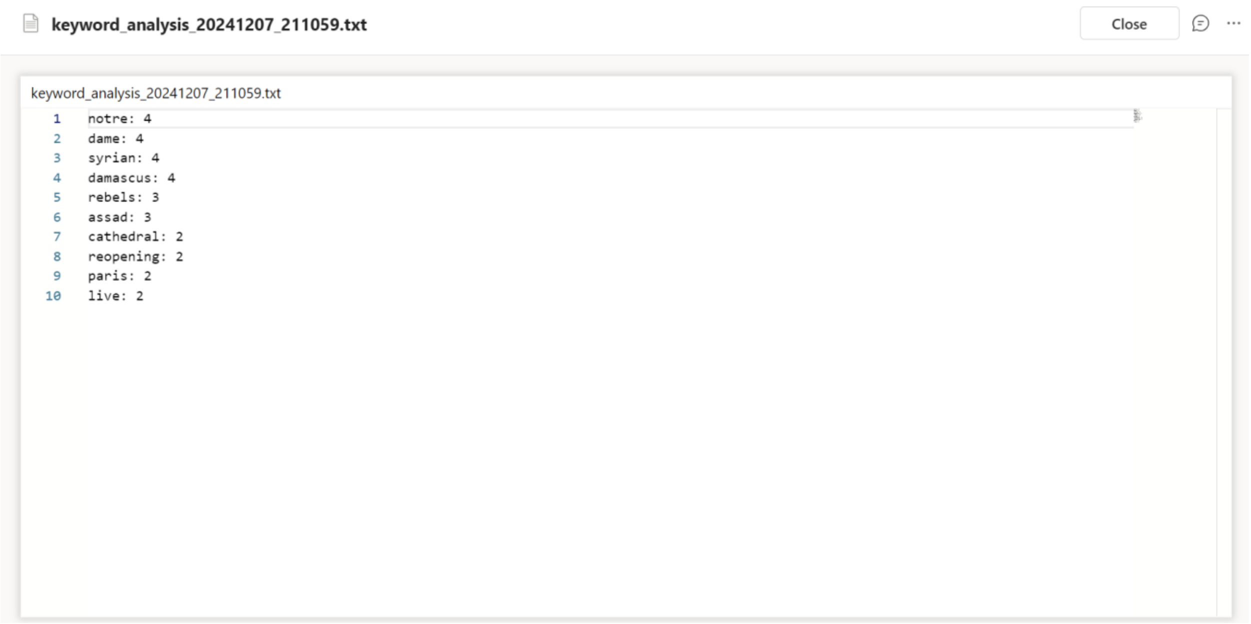
Visualizations

Detailed visualizations (e.g., word clouds, topic distributions) can be incorporated based on extracted data.

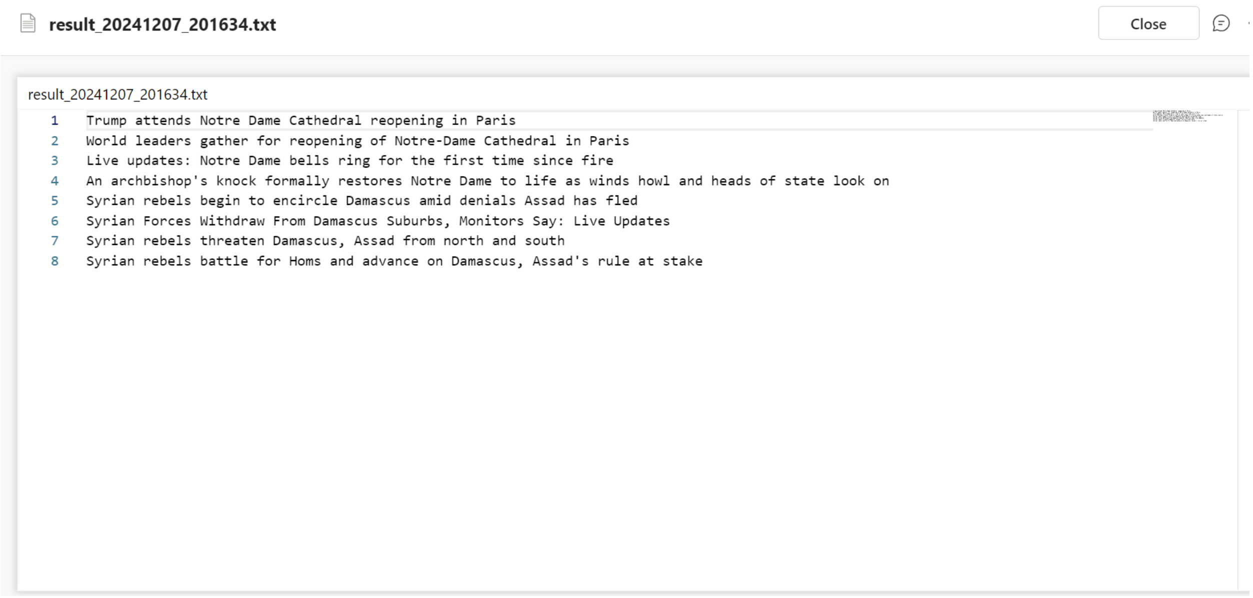


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	keywords/	Folder	-	-	-
□	results/	Folder	-	-	-


keyword_analysis_20241207_211059.txt
Close
keyword_analysis_20241207_211059.txt

```
notre: 4
dame: 4
syrian: 4
damascus: 4
rebels: 3
assad: 3
cathedral: 2
reopening: 2
paris: 2
live: 2
```



result_20241207_201634.txt

```
Trump attends Notre Dame Cathedral reopening in Paris
World leaders gather for reopening of Notre-Dame Cathedral in Paris
Live updates: Notre Dame bells ring for the first time since fire
An archbishop's knock formally restores Notre Dame to life as winds howl and heads of state look on
Syrian rebels begin to encircle Damascus amid denials Assad has fled
Syrian Forces Withdraw From Damascus Suburbs, Monitors Say: Live Updates
Syrian rebels threaten Damascus, Assad from north and south
Syrian rebels battle for Homs and advance on Damascus, Assad's rule at stake
```

Discussion

The pipeline successfully demonstrates the feasibility of automating news analysis. By focusing on trending topics, such as the Notre Dame reopening and the Syrian conflict, the system provides timely insights. However, future enhancements can include:

- Sentiment analysis for a deeper understanding of public opinion.
- Expanding data sources beyond Google News.

Conclusion

This project underscores the power of combining cloud-based tools with data science methodologies to address real-world challenges. The News Article Data Pipeline offers a scalable, efficient, and impactful solution for news analysis, with potential applications in journalism, academia, and beyond.

Contributions/References:

Martinaitis, D. (2020, June 23). *Serverless architecture for a web scraping solution | AWS architecture blog*. AWS Architecture Blog. https://aws.amazon.com/blogs/architecture/serverless-architecture-for-a-web-scraping-solution/

GeeksforGeeks. (2024, May 2). *Keyword extraction methods in NLP*. GeeksforGeeks. https://www.geeksforgeeks.org/keyword-extraction-methods-in-nlp/

***Python Requests Module*. W3Schools Online Web Tutorials. (n.d.). https://www.w3schools.com/python/module_requests.asp**