Intelligent News Summarization and Analysis

Overview

This project leverages various libraries and tools to fetch, process, analyze, and visualize news articles. It integrates with NewsAPI to retrieve articles based on specific queries, cleans the data for analysis, processes it using a large language model (LLM), and provides insights through clustering and trend analysis. The outcomes include summarization, sentiment analysis, and visualization of trends over time.

Components of the Code

1. Imports and Setup

The code starts by importing essential libraries for:

- HTTP requests (requests)
- Time handling (time)
- JSON processing (json)
- Text cleaning and web scraping (BeautifulSoup)
- AWS interaction (boto3)
- Data manipulation and analysis (numpy, pandas)
- Text vectorization and clustering (sklearn)
- Visualization (matplotlib, WordCloud)
- Environment variable management (dotenv)

2. Environment Variables

The project uses a .env file to securely load API keys needed for external services, specifically for accessing the NewsAPI and AWS.

```
load_dotenv()

api_key = os.getenv('API_KEY')
```

3. Data Collection

The NewsFetcher class is responsible for fetching news articles from the NewsAPI. It manages HTTP requests, including error handling for rate limits, and returns a list of articles based on the provided query.

```
class NewsFetcher:
  ...
  def fetch_news(self, query, page=1, page_size=10):
    ...
```

4. Article Processing

The ArticleProcessor class cleans the fetched articles, removing HTML tags and special characters, and extracts key information such as title, publication date, content, and source.

```
class ArticleProcessor:
  ...
  def process_articles(articles):
    ...
```

5. LLM Integration

The LLMProcessor class interacts with an AWS Bedrock model (Claude) to process articles. It generates summaries, extracts key points, assesses sentiment, and classifies topics based on the article content.

```
class LLMProcessor:
  ...
  def process_article(self, article):
    ...
```

Data:

	title	date	content	summary	key_points	sentiment	sentiment_score	topic	source.id	source.name
0	How Researchers Are Using Geospatial Technolog...	None	In 2014, after the disappearance of 43 Ayotzin...	Here is a concise summary of the key points fr...	[Here are the key points from the article:, , ...	To	0.2	Based on the content provided, the primary top...	wired	Wired
1	What Are Hall Effect Sensors and How Do They W...	None	Hall effect sensors are everywhere right now. ...	Here is a concise summary of the key points fr...	[Here are the key points from the article:, , ...	To	0.0	Based on the content provided, the primary top...	wired	Wired
2	Sonos announces 'breakthrough' Arc Ultra sound...	None	Sonos announces breakthrough Arc Ultra soundba...	Here is a concise summary of the article:\n\nS...	[Here are the key points from the article:, , ...	Based	0.7	Based on the content provided, the primary top...	the-verge	The Verge
3	Back to Its Prime Day Price (59% Off), The Rob...	None	The Roborock Q7 Max+ robot vacuum cleaner is c...	Here is a concise summary of the key points fr...	[Unfortunately, the provided text does not app...	To	0.9	Based on the content provided, the primary top...	None	Gizmodo.com
4	It Seemed Like an AI Crime-Fighting Super Tool...	None	In 2017, then 9-year-old Kayla Unbehaun was ab...	Here is a concise summary of the key points fr...	[Here are the key points from the article:, , ...	Based	-0.2	Based on the content provided, the primary top...	wired	Wired

```
{
  "articles": [
    {
      "title": "How Researchers Are Using Geospatial Technology to Uncover Mexico's Clandestine Graves",
      "date": null,
      "source": {
        "id": "wired",
        "name": "Wired"
      },
      "content": "In 2014, after the disappearance of 43 Ayotzinapa normalistas in Mexico, Silv\u00e1n and other CentroGeo professionals joined the scientific
advisory board on the case. During the search for the students,\u2026 [+2586 chars]",
      "summary": "Here is a concise summary of the key points from the article:\n\nThe article discusses the disappearance of 43 students from the Ayotzinapa
teachers' college in Mexico in 2014. After their disappearance, a group of professionals from CentroGeo, including Silv\u00e1n, joined the scientific advisory
board investigating the case. \n\nDuring the search efforts, CentroGeo used remote sensing data and analysis to identify possible clandestine graves and provi
de insights into the movements of the students based on cellphone data. Their work helped locate remains and guided search efforts.\n\nHowever, the article hi
ghlights disagreements between the CentroGeo team and the official investigation over the analysis and interpretation of evidence. CentroGeo's findings contra
dicted the government's claim that the students were killed and burned at a garbage dump.\n\nThe article portrays CentroGeo's role as providing independent, s
cientific analysis amid controversies surrounding the government's handling of the case. It emphasizes the importance of leveraging technology like remote sen
sing to search for the disappeared.\n\nUltimately, the full truth behind the students' disappearance remains unclear and a source of tension, with CentroGeo's
work offering an alternative perspective challenging the official narrative.",
      "key_points": [
        "Here are the key points from the article:",
        "",
        "1. In 2014, 43 student teachers (normalistas) from the Ayotzinapa Rural Teachers' College in Mexico disappeared.",
        "",
        "2. Geophysicists from the CentroGeo research group, including Jorge Silv\u00e1n, joined the scientific advisory board investigating the case.",
        "",
        "3. During the search for the students, CentroGeo used geophysical techniques like electrical resistivity tomography to look for clandestine graves.",
        "",
        "4. Their work helped locate remains and evidence related to the case in the town of Cocula.",
        "",
        "5. The Mexican government's initial account of the students' disappearance was contradicted by the scientific evidence found.",
        "",
        "6. CentroGeo's involvement highlighted the importance of independent scientific analysis in human rights cases.",
        "",
        "7. The case remains unresolved, with disagreements over what happened and where the students' remains are located.",
        "",
        "8. CentroGeo continues to assist in searching for clandestine graves related to the disappearance and other human rights cases in Mexico."
      ],
      "sentiment": "x"
```


6. Analysis and Insights

The InsightsAnalyzer class performs various analyses on the processed articles:

- **Clustering**: Uses TF-IDF vectorization and K-means clustering to group articles by topic.
- **Trend Analysis**: Analyzes trends over time based on article counts or specific keywords.
- **Sentiment Analysis**: Evaluates sentiment trends over time by calculating average sentiment scores.

```
class InsightsAnalyzer:

  ...

  def cluster_topics(self, articles):

    ...

  def trend_analysis(self, articles, keywords=None):

    ...

  def sentiment_trend_analysis(self, articles, sentiment_scores):

    ...
```

7. Visualization

The project includes methods for visualizing clusters and trends using Matplotlib, as well as generating a word cloud from the article content.

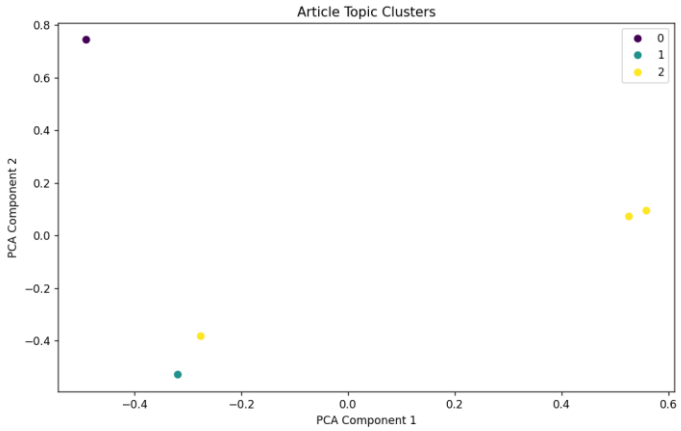
```
def plt_wordcloud(processed_article):

  ...
```

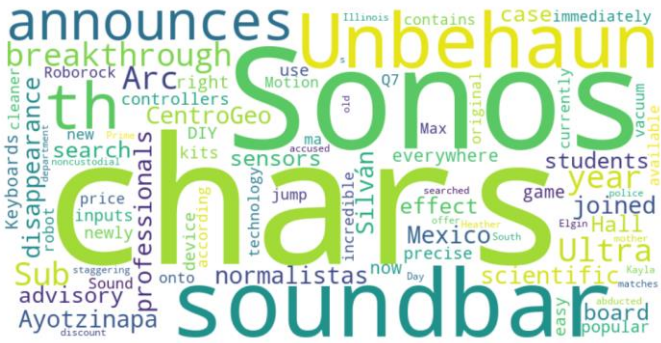
Workflow

1. **Fetching Data**:
    - The user initializes the NewsFetcher class with an API key and queries for news articles.
    - The fetch_news method retrieves articles related to the specified topic.
2. **Processing Articles**:
    - The articles are processed using the ArticleProcessor class to clean and extract relevant information.
3. **LLM Analysis**:
    - The LLMProcessor is utilized to summarize articles, identify key points, analyze sentiment, and classify topics.
4. **Insight Generation**:
    - The InsightsAnalyzer class clusters the articles and analyzes trends, allowing for keyword tracking and sentiment trends over time.
5. **Visualization**:
    - Visuals are generated for clusters and trends, helping to interpret the analysis effectively.
    - A word cloud is created from the article to visualize frequently occurring terms.

PCA Cluster:



Word Cloud:



Outcomes

The code produces several outcomes:

- A cleaned dataset of articles with titles, dates, sources, and processed content.
- Summaries, key points, and sentiment labels for each article.
- Clusters representing similar topics among articles.
- Trend analyses showing article counts or keyword occurrences over time.
- Visualizations for better understanding of data trends and article topics.

Conclusion

The project combines data collection, processing, and analysis, integrating modern NLP techniques with AWS capabilities. It provides a comprehensive solution for analyzing news articles, offering insights that can be used in various applications, including media monitoring, sentiment analysis, and content summarization.