Assignment 2: NLP Analysis of News Articles Using the NYTimes API

Assignment 2: NLP Analysis of News Articles Using the NYTimes API

Background

Natural Language Processing (NLP) plays a crucial role in extracting insights from large volumes of textual data. News articles contain valuable information about public sentiment, social trends, and emerging topics. Businesses, policymakers, and researchers use NLP to analyze trends in various domains, such as finance, politics, technology, and public health.

Industries utilize news analytics for:

- **Market Intelligence** – Companies monitor news for insights into competitors, industry trends, and economic conditions.
- **Public Sentiment & Brand Monitoring** – Organizations track public opinion about key figures (e.g., CEOs) or events.
- **Crisis Management** – Analyzing real-time news helps businesses manage PR crises and understand public perception.
- **Regulatory & Compliance Monitoring** – Financial and legal industries track regulatory changes and enforcement actions.
- **Risk Analysis** – Financial institutions and investors use news-based sentiment to predict market fluctuations.

This assignment will guide you through **collecting, processing, and analyzing news articles** to extract meaningful insights.

Project Tasks

Step 1: Build a Dataset Using the NYTimes API

- Use the **New York Times Article API** to collect articles based on:
    - **People Names** (e.g., Elon Musk, Jeff Bezos)
    - **Themes** (e.g., Gun Violence, Climate Change, AI advancements)
- Retrieve **articles from the past 2 years**.
- Ensure the dataset contains at least **50 rows**.
- For each article, extract:
    - **Title**
    - **Publication Date**
    - **Author**
    - **Location (if available)**
    - **Summary**
    - **Full Article Content**

Step 2: Save Data as a CSV File

- Store the collected dataset in a **CSV file**.
- This dataset must be submitted with your assignment.

Step 3: Load the Saved Dataset

- Read the dataset from the CSV file.
- Perform **basic inspection** (e.g., missing values, duplicates).

Step 4: Exploratory Data Analysis (EDA)

- **Feature Engineering:** Add numerical features to the text data, such as:
    - Word count per article
    - Sentence count
    - Character count
    - Average word length
- **Visualizations:** Create **multi-panel plots** to analyze: (Please use your own judgment)
    - **Article frequency over time** (e.g., monthly distribution)
    - **Distribution of article lengths**

- **Authors with the most articles**

Step 5: NLP Preprocessing Pipeline

- **Text Cleaning**: Remove **stop words, punctuation, special characters, and digits**.
- **Tokenization**: Use **SpaCy** for tokenizing the text into words.
- **Optional**: Use **lemmatization** or **stemming**.

Step 6: Convert Text Data Using TF-IDF

- Implement **TF-IDF Vectorization** to convert articles into numerical representation.

Step 7: NLP Application 1 – Sentiment Analysis

- Use **VADER sentiment analysis** to determine the sentiment of each article.
- **Analyze sentiment trends over time** to see how public perception changes.

Step 8: Word Cloud Analysis

- Generate **word clouds** for:
  - **Positive sentiment words**
  - **Negative sentiment words**

Step 9: NLP Application 2 – Topic Modeling

- **Use LDA (Latent Dirichlet Allocation)** to extract key topics.
- **Analyze topic trends over time**.

Part 2: Presentation (10-Slide Deck)

Your presentation should cover:

1. **Introduction** – Importance of news analytics & business value
2. **Data Collection** – Source, query parameters, dataset overview
3. **Exploratory Data Analysis** – Key statistics & visualizations
4. **Text Preprocessing** – Tokenization, cleaning, stopword removal
5. **Feature Engineering** – Word counts, sentence counts, TF-IDF
6. **Sentiment Analysis** – Trends & key findings
7. **Word Cloud Analysis** – Visual representation of positive & negative words
8. **Topic Modeling** – Key topics, trends, and insights
9. **Business Implications** – How businesses and policymakers use such insights

10. **Conclusion** – Key takeaways & future improvements

Submission Requirements

- **Jupyter Notebook (.ipynb)** – Well-documented code
- **Code notebook as pdf**
- **CSV dataset** – The dataset you created
- **PowerPoint Presentation (.pptx or Google Slides)** – No PDFs