Data Pipeline Summary

Overview

This document provides a concise summary of the data pipeline implemented for the VarsityPro Data Analyst/Data Systems Intern position. The pipeline involves web scraping, data cleaning, manipulation, visualization, and statistical analysis.

Approach

1. Retrieving Data:

- Utilized Python with BeautifulSoup for web scraping.
- Retrieved data from the Wikipedia page [Main_Page](https://en.wikipedia.org/wiki/Nikola_Tesla).

2. Extracting and Storing Data:

- Extracted paragraph text, plain text, and HTML text.
- Created Pandas DataFrames for text and HTML.
- Extracted hyperlinks and created a DataFrame.
- Stored DataFrames in text files and a CSV file.

3. Data Cleaning and Manipulation:

- Removed duplicate rows and handled missing values in hyperlinks DataFrame.

4. Data Visualization:

- Visualized text length distribution in 'Text' column using a histogram.
- Saved visualization as 'text_lengths_distribution.png'.

5. Statistical Analysis:

- Calculated summary statistics for the 'Link' column using the `describe()` method.

Tools/Libraries Used

- Python: Scripting and data manipulation.
- BeautifulSoup: Web scraping.
- Pandas: Data manipulation and analysis.
- Matplotlib: Data visualization.

Challenges Faced

- Web Scraping Challenges: Dealing with HTML structure and dynamic content.
- Data Cleaning: Deciding on handling missing values and outliers.
- Tool Selection: Choosing the right combination of libraries.

Conclusion

The data pipeline successfully retrieves, cleans, manipulates, visualizes, and analyzes data from a Wikipedia page. It demonstrates proficiency in Python, web scraping, data manipulation, and visualization.

For detailed information, refer to the script `Wikipedia scraping.ipynb` and the accompanying documentation.