DAY1

- Task 1: Git, GitHub, and Initial Data Analysis
    - Set up a Python environment and a GitHub repository named news_correlation_10ac_week0.
    - Create branches for daily tasks starting with task-1.
    - Modify existing codebase to shift focus from slack data to news data (e.g., rename SlackDataLoader to NewsDataLoader).
    - Conduct exploratory data analysis (EDA) and statistical analysis on news data to answer specific questions about news websites, traffic, and content related to various regions and sentiments.
    - Commit changes frequently with descriptive messages.

Insights:

- ✓ ETFDaily, Globenewswire and Globalsecurity.org are the websites with the largest number of news articles.
- ✓ All Africa, CNN, The verge are the websites with the smallest number of articles.
- ✓ ETFdaily, globalnewswire and global security org have the highest number of visitors traffic
- ✓ Allafrica, bbc and the verge have the least number of visitors traffic.
- ✓ US, UK, Italy is where most media companies originate from.
- ✓ Most articles mention US followed by Canada.
- ✓ Globenewswire seems to be a site that mentions US, UK, Africa, Middle East relatively frequently.
- ✓ AlJazeera mentions Middle East in its articles quite a lot.

- ✓ Distribution of message length by website domain does not vary much, however there are some outliers.

DAY2

- Task 2: Data Science Component Building
    - From the main branch, create a new branch task-2 to further develop analysis capabilities.
    - Introduce MLOps components such as feature store, model versioning, and CI implementation using tools like Docker and GitHub Actions.
    - Perform advanced analyses like keyword extraction, topic modeling, and sentiment analysis.
    - Develop predictive models and possibly network analysis to understand the relationships and trends within the news data.
    - Summarize different MLOps components and their applications.

Insights:

- ✓ The top 5 keywords are: chars, report, free, 2023 and according.
- ✓ An average cosine similarity score of 0.026437723399909175, indicates a very low degree of similarity between the vectors being compared.
- ✓ Keywords in the headline/title are not that similar compared to keywords in the news body across sites

✓ The top words in each topic are:`

- 0: ['chars', 'gaza', 'israel', '2023', 'said', 'people', 'israeli', 'hamas', 'november', 'war'],
- 1: ['according', 'recent', 'chars', 'quarter', 'report', 'free', 'filing', 'company', 'second', 'shares'],
- 2: ['free', 'report', 'chars', 'shares', 'reports', 'quarter', 'holdings', 'owned', 'firm', 'fund'],
- 3: ['2023', 'nov', 'chars', 'globe', 'newswire', 'market', 'global', 'billion', 'report', 'company'],
- 4: ['earnings', 'chars', 'company', 'report', 'free', 'reported', 'share', 'quarter', 'results', 'eps'],
- 5: ['november', 'chars', 'report', 'dividend', 'free', 'record', 'reports', 'transaction', 'thursday', 'announced'],
- 6: ['chars', 'new', 'getty', 'world', 'images', 'years', 'year', 'time', 'just', 'companies'],
- 7: ['report', 'free', 'chars', 'research', 'rating', 'issued', 'reports', 'shares', 'morning', 'price'],
- 8: ['traded', 'stock', 'shares', 'free', 'chars', 'report', 'trading', 'high', 'price', 'low'],
- 9: ['chars', 'october', 'president', '2023', 'november', 'state', 'minister', 'short', 'said', 'prime

DAY3

- Task 3: Database Integration
    - After merging previous tasks, create a new branch task-3 to focus on database integration.
    - Design and implement a PostgreSQL database schema to store machine learning features.
    - Load data into PostgreSQL, ensuring it is structured to support the ML models developed in Task 2.

Insights:

- ✓ Databases and schemas are really critical. Postgres sql is also a heavy technology
- ✓ You can use databases with python.
- ✓ Depending on the type of schema, normalization might be necessary.