10 Academy

Week 0 - Report

Name: Yishak Tadele

Report One

Understanding

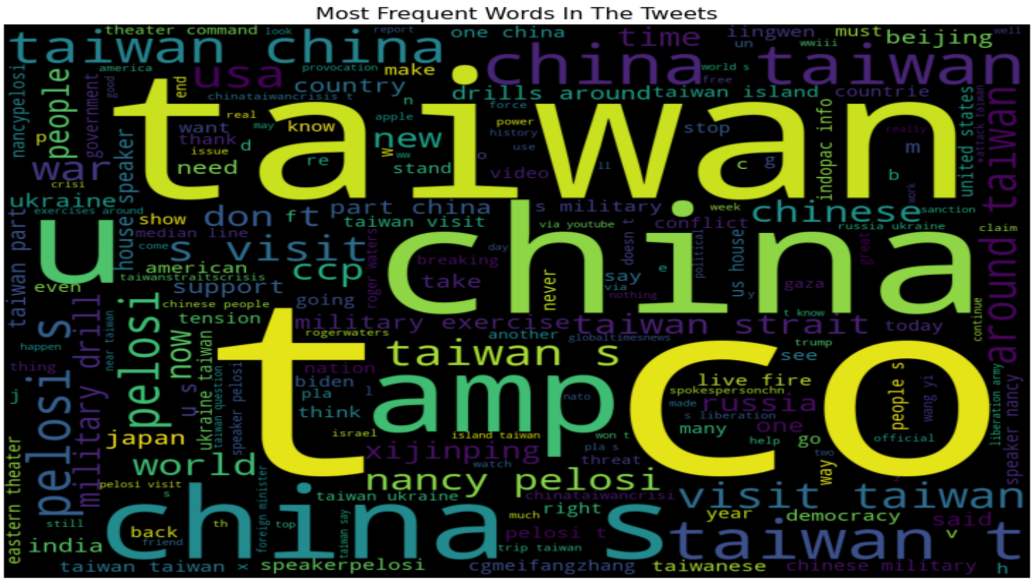
The week 0 project is a Machine Learning project based on a Twitter dataset collected using keywords that reflect issues around the relations or interactions of the two countries China and the USA.

The dataset consists of tweets and general information about the tweet status with similar hashtags.

- The main objective of the project is to analyse the data to find patterns that will be able to express the fate of the relations between the two countries.
- From the technical side, in the project, we will be able to develop ML skills such as
    - Fetching and Extracting data
    - Explorative analysis and data visualization
    - Experiment with ML models and Evaluation metrics
    - ML workflow pipelines and etc

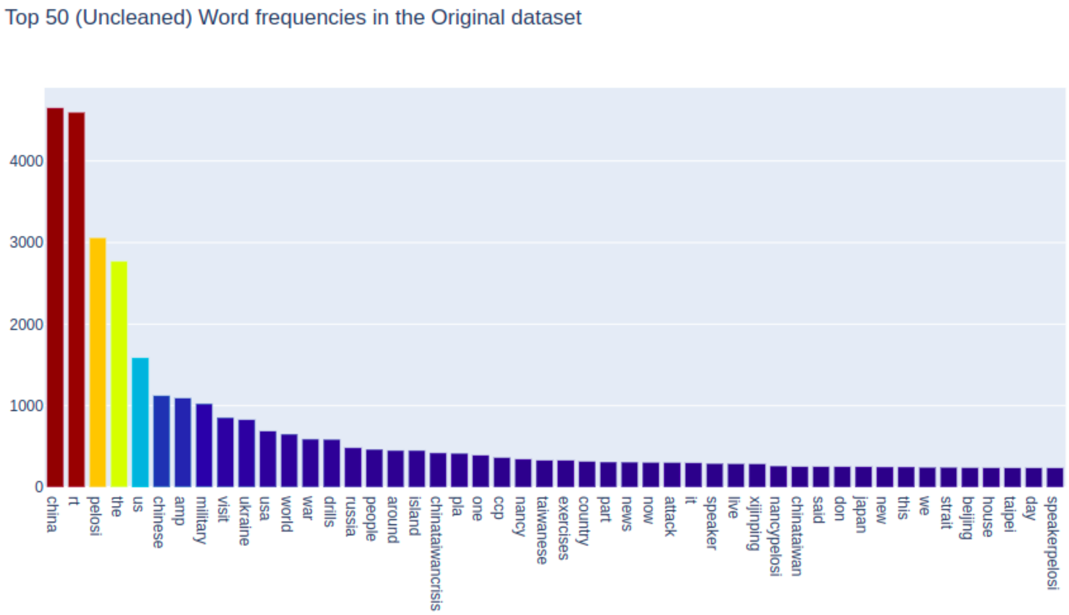
Progress

1. Github Setup: I was able to create my git repo from the template provided and perform the tasks that are stated on the readme file.
2. Github Actions for CI: I am able to create a GitHub action that runs a program to test the code for any push/pull request.
3. Data Analysis: I was trying to get a deeper insight into the dataset so I have created a Data Exploration notebook file and tried some methods.
    - In the notebook, I have tried to clean Twitter and plotted the most frequent words as follows

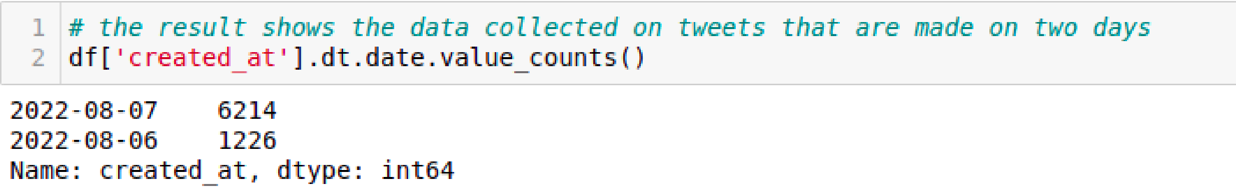


Most Frequent Words In The Tweets

- Also, the top words are plotted as follows.



- I was also able to find out that the global_twitter dataset is collected in two days, and can be found in the notebook.



```
# the result shows the data collected on tweets that are made on two days
df['created_at'].dt.date.value_counts()
```

```
2022-08-07    6214
2022-08-06    1226
Name: created_at, dtype: int64
```

- The dataset consists of a single language, which is English. This and other findings are shown in the notebook I provided in my EDA branch.