- **Name of the group members and CNetIDs**
  Siwen Chen (chens1)
  Wenxin Gu (wguab)
  Carolyn Liu (crliu)

- **A brief overview of the final project (200 words maximum)**
  Stock prices are influenced by the public attitude towards the market. Psychological research shows that emotions play a significant role in human decision-making. Similarly, behavioral finance research shows that financial decisions are influenced by people's emotions and moods.

  Twitter is a popular place for users to express their moods and sentiments towards a variety of topics. We look into how the polarity of sentiments in tweets is related to the daily movements of stock prices!

  Data:

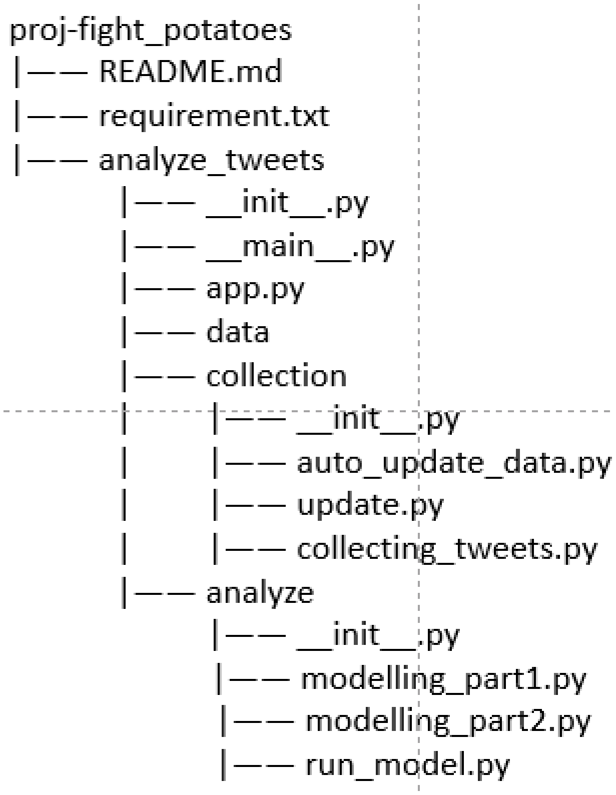
  Tweets were collected from the Twitter Developer API. We gathered around 1.6 million tweets related to COVID-19 and US-China Relations from Feb 14 to March 15.

  The query keywords used to collect tweets are as follows:

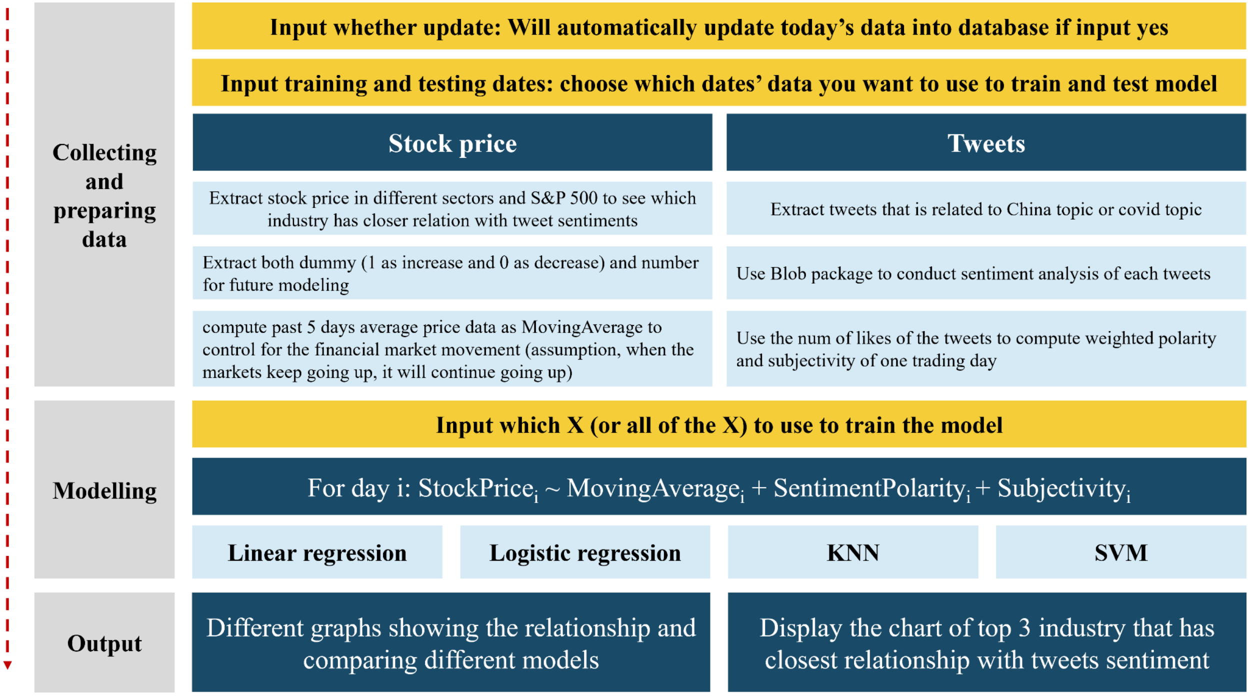
  > COVID-19: covid19, covid, covid-19, vaccine, vaccination, omicron, booster, mask, mandate, lockdown, death rate, travel restriction, total infections, breakthrough infections, social distancing, quarantine, isolation, pandemic, & shutdown.
  >
  > US-China Relations: Trump, Biden, trade war, Xi Jinping, Eileen Gu, Taiwan, Hong Kong, CCP, TikTok, Huawei, tariffs, human rights, Xinjiang, Zhao Lijian, & Ned Price.

  We collected stock data from all S&P 500 companies and grouped the data into 11 different sectors: Information Technology, Health Care, Consumer Discretionary, Financials, Communication Services, Industrials, Consumer Staples, Energy, Real Estate, Materials, and Utilities.

- **The overall structure of the software (1-page maximum). It would be nice to include a helpful diagram of how the modules are connected with each other but this is not required.**



```
proj-fight_potatoes
|—— README.md
|—— requirement.txt
|—— analyze_tweets
        |—— __init__.py
        |—— __main__.py
        |—— app.py
        |—— data
        |—— collection
        |       |—— __init__.py
        |       |—— auto_update_data.py
        |       |—— update.py
        |       |—— collecting_tweets.py
        |—— analyze
                |—— __init__.py
                |—— modelling_part1.py
                |—— modelling_part2.py
                |—— run_model.py
```



Collecting and preparing data

Input whether update: Will automatically update today’s data into database if input yes

Input training and testing dates: choose which dates’ data you want to use to train and test model

Stock price

Extract stock price in different sectors and S&P 500 to see which industry has closer relation with tweet sentiments

Extract both dummy (1 as increase and 0 as decrease) and number for future modeling

compute past 5 days average price data as MovingAverage to control for the financial market movement (assumption, when the markets keep going up, it will continue going up)

Tweets

Extract tweets that is related to China topic or covid topic

Use Blob package to conduct sentiment analysis of each tweets

Use the num of likes of the tweets to compute weighted polarity and subjectivity of one trading day

Modelling

Input which X (or all of the X) to use to train the model

For day i: $\text{StockPrice}_i \sim \text{MovingAverage}_i + \text{SentimentPolarity}_i + \text{Subjectivity}_i$

Linear regression

Logistic regression

KNN

SVM

Output

Different graphs showing the relationship and comparing different models

Display the chart of top 3 industry that has closest relationship with tweets sentiment

- **A description on the code responsibilities for each group member (i.e., who was responsible for what module, files, tasks, etc.).**

  The responsibilities for completing this project consist of three parts, including the data gathering and cleaning, the data processing and analysis, and the data visualization part.

  We worked on the three parts together and were individually responsible for small tasks, specifically:

  - Data gathering and cleaning:

    1. Read the existing paper related to our topic (Siwen, Wenxin, Carolyn)
    2. Use Twitter API to collect tweets about US-China relationship and Covid (Wenxin, Carolyn, Siwen take turns every weekday)
    3. Collect the financial data using yfinance package (Carolyn, Siwen, Wenxin)
    4. Make the app auto-update the data of today (Wenxin)

  - Data processing and analysis:

    1. Clean and prepare the dataset including using Textblob package to perform sentiment analysis, extracting Xs and Y from the datafile (Wenxin)
    2. Train a positve_negative_determination model using the training dataset using the different models’ results as classifiers including linear regression, logistic regression, KNN, and SVM.
    (Carolyn, Siwen, Wenxin)

- Data visualization/output:
  Use Dash to create the layout of the web application
  1. Dropdowns (Carolyn)
  2. Graphs for visualizing the output (Carolyn)
  3. The top 3 sectors result for different models (Siwen)
  4. The tabs explaining the data and output (Siwen)

- **Short description on how to interact with the application and what it produces.**

  Users can interact with the project through the command line and the web application. Screenshots of the output are presented in README.md.

  There are 3 user input opportunities in the command line:

  (1) users can choose training and testing dates used in the models;
  (2) users can choose to have the tweets and the financial data auto-update for the day you interact with this project;
  (3) finally, users are able to choose different independent variables (polarity of the tweets, subjectivity of the tweets and/or 5-day moving average)
  All command line inputs should be integers representing dates/variables; multiple selections should be separated by a space (examples are given in the command line prompt).

  In the web application, the application will produce two graphs and one table. One of the graphs is for the testing data and the other one is for training data. Users can check out the different graphs through the various dropdown menus.

  You can choose a topic (Covid or China) and a model to see the accuracy of the prediction using average Twitter sentiments on stock market movement. If you choose the linear regression model, you can also select a sector. For all other models, we show the accuracy of the models with all sectors on one plot.

- **What the project tried to accomplish and what it actually accomplished (200 words)**

  The project tried to analyze how sentiments in tweets about US-China relations and COVID-19 affect movements in the stock market.

  We trained several simple models (linear/logistic regression, KNN, & SVM) with added controls such as 5-day moving averages and subjectivity of the tweets, but there are many more factors that affect the stock market that we weren’t able to include. We were also limited by the number of tweets we could collect (a quota of 1.5m tweets/month among the group members); we managed to collect a month’s worth of tweets. Due to the limitations on the control variables and the short time frame, our accuracy rate is not as high as we would have liked it to be. Additionally, we give users the opportunity to add more testing data with an auto-update feature of the application, which collects tweets and stock data from the day the application is accessed (assuming a day when the markets are open).