

User Manual

# Data Mining News and Social Media for the Fertile Ground Project

---

Rishabh Shrestha

Ting Xiao

Dr. Dawn Wilkins



## Introduction:


Data Mining is the process of extracting and discovering information from raw data gathered online or through various ways. It involves finding patterns in large datasets involving methods at the intersection of machine learning, statistics, and database systems.

In this project, we extract data related to the fertile ground project in Jackson, MS from online sources and then perform meaningful analysis and visualization to understand the change in frequency and nature of content among the public about the topics related to food access, food security, and fertile ground project. The data is collected from social media sources such as Twitter using an API key and scraping library in Python and similarly, using google news API for news sources. After the collection of data, datasets are built to perform visualization, sentiment analysis. Machine Learning models are also built to predict the sentiment as positive or negative. The model is built after training it with an online tweet dataset since we were not able to collect enough data to train and test the Machine Learning model to predict the sentiment.

This document, specifically, is the user manual of this project. The document will explain the various functionalities of the project as well as how to run and test the project. The document will walk you through the necessary installations, system and software requirements, required packages and libraries, and the scripts of the program written in Python and then run the code.

## System and Installation Requirements:

1. System Requirements:
  - a. This project was developed with Python version 3.8.7 so the Python version of 3.8 or above is recommended. The development environment of the project was Jupyter Notebook. The installation of the Anaconda is highly recommended and then use that Anaconda distribution to install Python and Jupyter. Download Anaconda from the official website [anaconda.com](https://anaconda.com) and be sure to download Anaconda's latest Python 3 version. Follow the instructions on the official website about downloading and installing. After installing that you can run "*jupyter notebook*" on your terminal/ powershell and then the jupyter notebook will start on your local server. Be sure to add Path in your Environment Variables. After you download Python, set your Path to Scripts folder inside the Python folder. Similarly, if



you download Anaconda, be sure to set the path to Scripts folder inside Anaconda folder and to the bin folder inside Library folder of Anaconda.

- b. Similarly, the above-mentioned steps can be done, especially for experienced Python users i.e. Installing Jupyter with pip. As an existing Python user, you may wish to install Jupyter using Python's package manager, pip, instead of Anaconda. First, ensure that you have the latest pip; older versions may have trouble with some dependencies and then run *"pip install jupyter"* in your terminal or command line. After you install, you can just run "jupyter notebook " to run the notebook on your local server.
- c. For device and OS specification, this project was developed on Windows 10 OS with intel core i5 processor and 8 GB RAM so any devices with that specification or higher will be able to run the project. Running the project on macOS won't have any issues either.
- d. This system requires the local server to function as a web server to open jupyter notebook and python files.

## 2. Required packages:

- a. NumPy
- b. Pandas
- c. Scikit Learn
- d. Tweepy
- e. Json
- f. Datetime
- g. Csv
- h. String
- i. Preprocessor
- j. Os
- k. Time
- l. Snsrape
- m. GoogleNews
- n. Newspaper3k
- o. Nltk
- p. vaderSentiment
- q. Matplotlib
- r. Seaborn
- s. Scipy

- t. Re
- u. Python-dateutil
- v. Plotly
- w. Collections
- x. Palettable
- y. Random
- z. Dotenv
- aa. Wordcloud
- bb. Datanews
- cc. Math
- dd. Glob
- ee. Requests
- ff. BeautifulSoup
- gg. pandas\_dedupe

All of these packages can be installed using the command “pip install <package-name>”.  
For example, *pip install pandas*.

#### Required python scripts:

The whole project is available as a GitHub repository on this link-

[https://github.com/T-Xiao/CERE\\_fertile\\_ground](https://github.com/T-Xiao/CERE_fertile_ground)

1. The python script for scraping tweets then creating csv files can be found under this link as tweetScape.ipynb file.

[https://github.com/T-Xiao/CERE\\_fertile\\_ground/blob/main/tweetScape.ipynb](https://github.com/T-Xiao/CERE_fertile_ground/blob/main/tweetScape.ipynb)

2. The python script for scraping news articles from Datanews API with detailed steps can be accessed and downloaded using this link as DatanewsAPI.ipynb file.

[https://github.com/T-Xiao/CERE\\_fertile\\_ground/blob/main/DatanewsAPI.ipynb](https://github.com/T-Xiao/CERE_fertile_ground/blob/main/DatanewsAPI.ipynb)

3. The python script for scraping tweets then creating csv files can be found under this link as googleNewsScape.ipynb

[https://github.com/T-Xiao/CERE\\_fertile\\_ground/blob/main/googleNewsScape.ipynb](https://github.com/T-Xiao/CERE_fertile_ground/blob/main/googleNewsScape.ipynb)

4. The python script to perform visualizations and analysis of tweets can be found here as tweetAnalysis.ipynb file

[https://github.com/T-Xiao/CERE\\_fertile\\_ground/blob/main/tweetAnalysis.ipynb](https://github.com/T-Xiao/CERE_fertile_ground/blob/main/tweetAnalysis.ipynb)

5. The python script to perform Machine Learning to train and test models to predict the sentiment of tweets can be found here as [machineLearning.ipynb](#) file

[https://github.com/T-Xiao/CERE\\_fertile\\_ground/blob/main/machineLearning.ipynb](https://github.com/T-Xiao/CERE_fertile_ground/blob/main/machineLearning.ipynb)

### **Usage of the Project:**

A. The main user of this project will be the client- Center for Research Evaluation(CERE) at Ole Miss. Follow the instructions below to run the project and get the results.

#### How to run the scripts:

At the very first, clone the GitHub repository or download the folder of the project from the Github link of the repository provided above. After you have download the project follow the instructions below:

First, run the jupyter notebook in the terminal or command line. To do that, go to the project folder in the terminal. Now, to run the jupyter notebook, just type `'jupyter notebook'` in the terminal and hit enter. This should run the local server of the jupyter notebook in your browser. After that you will see the list of files inside the project folder including scripts and .csv files.

1. Go to `datascape.ipynb` file to scrape the tweets and news articles. After you are in the file run the whole file by clicking Cell in the menu of the notebook file and select Run All. Instead, you can also run each cell and hit *Shift+Enter* to run one cell at a time and then hit *Shift+Enter* again to run the cell below it and so on. After you finish running all the cells, you can check for the CSV files created from the scraped data in the project folder. The name of the CSV files will depend upon the type of data you will scrape. For that, refer to the particular cells at the end of the cell to check for the name of the .csv files that will be created after scraping.

Here is a sample of tweet scrape after you run the cell. The cells will have a header for what each of them do, so it will be easy to understand. You can see

what kind of fields you will have access to for each tweet in the dataframe below. The dataframe below will be stored in a CSV file and downloaded following this.

```

for i,tweet in enumerate(sntwitter.TwitterSearchScraper(term+' geocode:32.2988,-90.1848,50km since:2018-03-01').get_items()):
    if i>500:
        break

    tweets_list2.append([tweet.date, tweet.id, tweet.content, tweet.user.username, tweet.user.location, tweet.replyCount, tweet.retweetCount, tweet.likeCount, tweet.quoteCount])

tweets_df2 = pd.DataFrame(tweets_list2, columns=['Datetime', 'Tweet Id', 'Text', 'Username', 'User Location', 'Reply Count', 'Retweet Count', 'Like Count', 'Quote Count'])
tweets_df2.drop_duplicates(subset="Tweet Id", keep=False, inplace=True)
tweets_df2

```

Out[4]:

	Datetime	Tweet Id	Text	Username	User Location	Reply Count	Retweet Count	Like Count	Quote Count
0	2021-01-22 04:12:06+00:00	1352469068452868103	@GanuchauAdam Thank you for retweeting! Also ...	_fertileground_	Jackson, MS	0	0	0	0
1	2021-01-22 01:00:14+00:00	1352420775530291202	We've BEEN asking for this!nKey House leade...	_fertileground_	Jackson, MS	0	0	0	0
2	2020-11-30 22:02:50+00:00	1333531965245566977	Mississippi Goals! https://t.co/DiQEOYcn3j	_fertileground_	Jackson, MS	0	0	2	0
3	2020-11-30 22:01:50+00:00	1333531710907150336	Tips! https://t.co/lyte1eETRN	_fertileground_	Jackson, MS	0	0	0	0
4	2020-11-18 12:44:34+00:00	1329042818199908353	@JacksonStateU Athletics and @UHC will be prov...	_fertileground_	Jackson, MS	0	0	0	0
...	...	...	...	...	...	...	...	...	...

Similarly, here is a sample of news article scrape after you run the cell pertaining to it. You will get access to fields like Date Published, Media Source, Title of Article, Article Content, Summary of the article. The dataframe below will be stored in a CSV file and downloaded following this.

```

caught a timeout
caught a timeout

```

C:\Users\rishabhstha\AppData\Local\Programs\Python\Python38\lib\site-packages\dateutil\parser\\_parser.py:1213: UnknownTimezoneWarning: tzname EDT identified but not understood. Pass "tzinfos" argument in order to correctly return a timezone-aware datetime. In a future version, this will raise an exception.

warnings.warn("tzname {tzname} identified but not understood. ")

Out[9]:

	Date	Media	Title	Article	Summary
0	Oct 10, 2020	Mississippi State University	MSU's Keenum applauds selection of World Food ...	Contact: James Carskadon\n\nMSU President Mark...	(Submitted Photo)STARKVILLE, Miss — Mississipp...
1	Dec 17, 2020	WJTV	Mississippi Food Network helps with food inse...	JACKSON, Miss. (WJTV) – The coronavirus pandem...	(WJTV) – The coronavirus pandemic has impacted...
2	Feb 26, 2020	Jackson Free Press	Jackson's Food Insecurity Focus in April Expo	In a state where one in four children go to be...	The U.S. Department of Agriculture defines foo...
3	Nov 23, 2020		Survey: Food Insecurity, Lack of Support Remai...	JACKSON, Mississippi - Since the start of the ...	JACKSON, Mississippi - Since the start of the ...
4	May 4, 2018	Mississippi Today	Mississippi still the hungriest state	For the eighth straight year, Mississippi has ...	This is the eighth edition of Feeding America'...
...	...	...	...	...	...
128	Aug 21, 2020	WCBI	Food insecurity remains a concern for many fam...	During this time of uncertainty, dozens of fam...	During this time of uncertainty, dozens of fam...

- Now, to perform visualizations and analysis, run the dataAnalysis.ipynb file in the project folder the same way as above and you will see the visualizations and analysis performed pertaining to the particular cell. You will be able to see visualizations such as time-series graphs of counts of tweets and counts of news articles. Similarly, you will be able to see comparison of different sentiment scores of tweets and news articles and other graphs related to it. There are many other

visualizations that you will be able to see after you run the notebook. The notebook will have a header in every cell describing what each cell will produce.

### Most Common Words




The image above is one of interactive visualizations. It shows the most common words in tweets after preprocessing and cleaning. Once you hover over the words, it will show the count of the word in the tweet dataset.

3. Similarly, to build the Machine Learning model and see the comparisons of the performance of the different models, run the machineLearning.ipynb file the same way too.

### Uninstall/Removal Process:

To stop running the jupyter notebook , you can shut down the kernel by hitting Ctrl+C in the terminal where your jupyter notebook is hosted from and it will show the message for



kernel interruption and shutdown. The temporary stack memory will be cleared and you will have to load and run the files again the next time you want to see the output.

To delete the project from your system, you can remove all the files pertaining to the project and project folder itself by deleting them. This will either remove them from the system or move them to a “trash bin” where the files may be permanently removed.

### **Administration and Maintenance:**

The collection of data and analysis for the Fertile Ground Project will be administered even after this senior project by the clients(Center for Research Evaluation Ole Miss). The collection of data for specific periods, locations, and keywords will be done accordingly as per need. The users will perform the query as per their need to scrape the data and perform analysis.

### **Enhancement and Future Ideas:**

In the future, more data could be collected for a longer time period to get more insights from the data. A sentiment lexicon for the level of enthusiasm of online posts could be created. Similarly, scraping the data from Facebook and Instagram by building an API could be done to get more data. Machine Learning on the collected data to train and test models could be done once we have access to more data.

Note- You can reach out to me at [rishabhstha@gmail.com](mailto:rishabhstha@gmail.com) if you have any questions or run into any issues regarding the project.