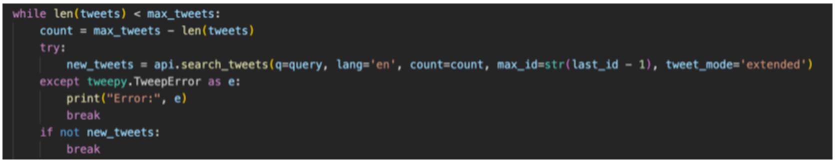
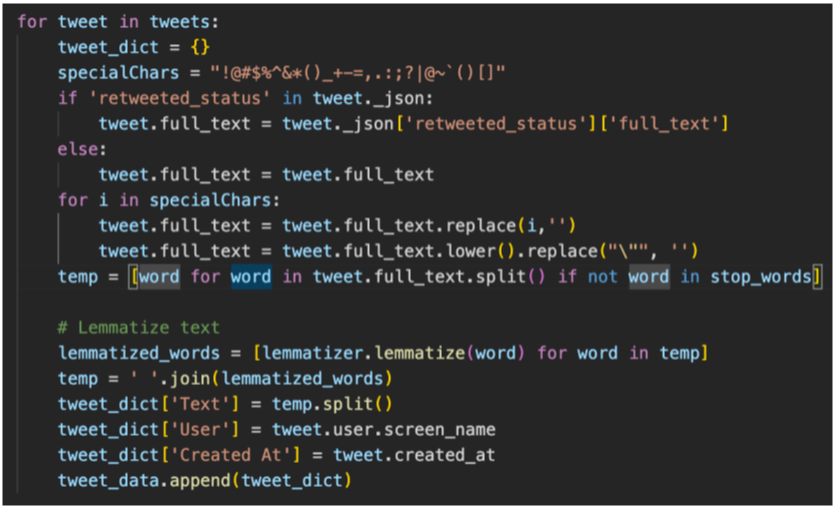
Twitter Sentiment Detection for ChatGPT

In the project we use creepy to collect data. Tweepy is a Python library for accessing the Twitter API. It provides a convenient way to use the Twitter API to interact with Twitter data, including retrieving tweets, user information, and more. We use the keyword “chatgpt” to collect 20000 datasets and preprocess and EDA. Here is the code for collecting data.



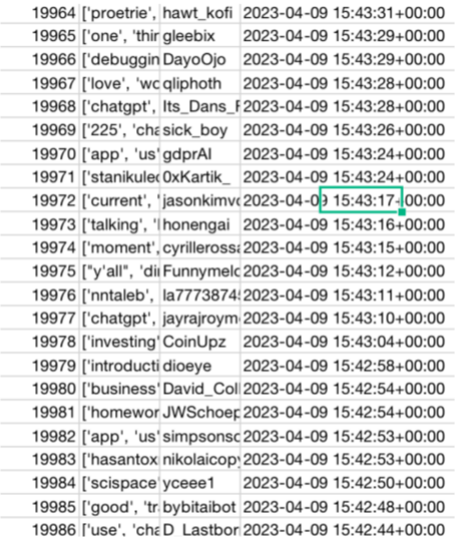
```python
while len(tweets) < max_tweets:
    count = max_tweets - len(tweets)
    try:
        new_tweets = api.search_tweets(q=query, lang='en', count=count, max_id=str(last_id - 1), tweet_mode='extended')
    except tweepy.TweepError as e:
        print("Error:", e)
        break
    if not new_tweets:
        break
```

Here is the code for preprocessing.



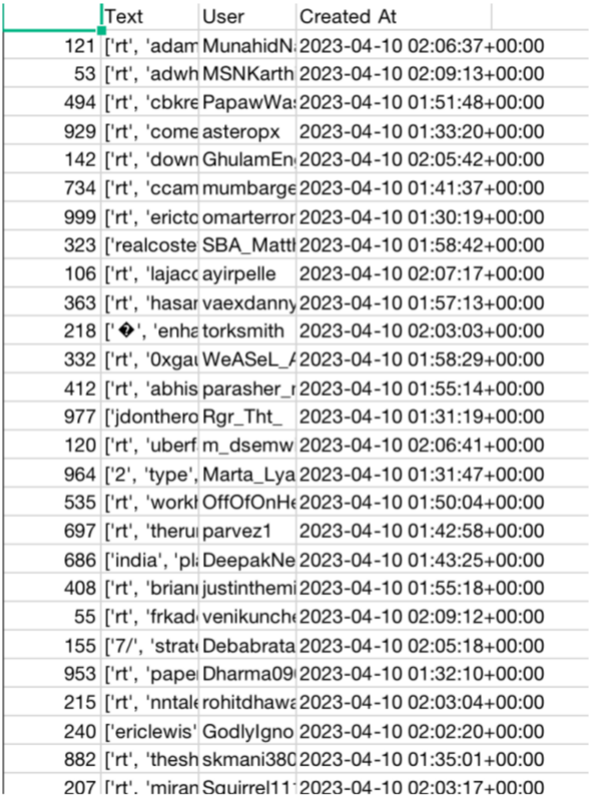
```python
for tweet in tweets:
    tweet_dict = {}
    specialChars = "!@#$%^&*()_+-=,.:;?|@~`()[]"
    if 'retweeted_status' in tweet._json:
        tweet.full_text = tweet._json['retweeted_status']['full_text']
    else:
        tweet.full_text = tweet.full_text
    for i in specialChars:
        tweet.full_text = tweet.full_text.replace(i,'')
        tweet.full_text = tweet.full_text.lower().replace("\"", '')
    temp = [word for word in tweet.full_text.split() if not word in stop_words]

    # Lemmatize text
    lemmatized_words = [lemmatizer.lemmatize(word) for word in temp]
    temp = ' '.join(lemmatized_words)
    tweet_dict['Text'] = temp.split()
    tweet_dict['User'] = tweet.user.screen_name
    tweet_dict['Created At'] = tweet.created_at
    tweet_data.append(tweet_dict)
```

When all the data collected. We create two files to save the data. One is tweets.csv which including 20000 datasets. The other is sample.csv which including 1000 datasets. Here is tweets.csv



| | | | |
|---|---|---|---|
| 19964 | ['proetrie', | hawt_kofi | 2023-04-09 15:43:31+00:00 |
| 19965 | ['one', 'thir | gleebix | 2023-04-09 15:43:29+00:00 |
| 19966 | ['debuggin | DayoOjo | 2023-04-09 15:43:29+00:00 |
| 19967 | ['love', 'wc | qliphoth | 2023-04-09 15:43:28+00:00 |
| 19968 | ['chatgpt', | Its_Dans_F | 2023-04-09 15:43:28+00:00 |
| 19969 | ['225', 'cha | sick_boy | 2023-04-09 15:43:26+00:00 |
| 19970 | ['app', 'us' | gdprAI | 2023-04-09 15:43:24+00:00 |
| 19971 | ['stanikulec | 0xKartik_ | 2023-04-09 15:43:24+00:00 |
| 19972 | ['current', ' | jasonkimvc | 2023-04-09 15:43:17+00:00 |
| 19973 | ['talking', 'I | honengai | 2023-04-09 15:43:16+00:00 |
| 19974 | ['moment', | cyrillerossa | 2023-04-09 15:43:15+00:00 |
| 19975 | ["y'all", 'dii | Funnymelc | 2023-04-09 15:43:12+00:00 |
| 19976 | ['nntaleb', | la77738745 | 2023-04-09 15:43:11+00:00 |
| 19977 | ['chatgpt', | jayrajroym | 2023-04-09 15:43:10+00:00 |
| 19978 | ['investing' | CoinUpz | 2023-04-09 15:43:04+00:00 |
| 19979 | ['introducti | dioeye | 2023-04-09 15:42:58+00:00 |
| 19980 | ['business' | David_Coll | 2023-04-09 15:42:54+00:00 |
| 19981 | ['homewor | JWSchoep | 2023-04-09 15:42:54+00:00 |
| 19982 | ['app', 'us' | simpsonsc | 2023-04-09 15:42:53+00:00 |
| 19983 | ['hasantox | nikolaicopy | 2023-04-09 15:42:53+00:00 |
| 19984 | ['scispace' | yceee1 | 2023-04-09 15:42:50+00:00 |
| 19985 | ['good', 'tr | bybitaibot | 2023-04-09 15:42:48+00:00 |
| 19986 | ['use', 'cha | D_Lastbor | 2023-04-09 15:42:44+00:00 |

Here is sample.csv



| | Text | User | Created At |
|---|---|---|---|
| 121 | ['rt', 'adam | MunahidNa | 2023-04-10 02:06:37+00:00 |
| 53 | ['rt', 'adwh | MSNKarth | 2023-04-10 02:09:13+00:00 |
| 494 | ['rt', 'cbkre | PapawWa: | 2023-04-10 01:51:48+00:00 |
| 929 | ['rt', 'come | asteropx | 2023-04-10 01:33:20+00:00 |
| 142 | ['rt', 'down | GhulamEn | 2023-04-10 02:05:42+00:00 |
| 734 | ['rt', 'ccam | mumbarge | 2023-04-10 01:41:37+00:00 |
| 999 | ['rt', 'erictc | omarterror | 2023-04-10 01:30:19+00:00 |
| 323 | ['realcoste | SBA_Matth | 2023-04-10 01:58:42+00:00 |
| 106 | ['rt', 'lajacc | ayirpelle | 2023-04-10 02:07:17+00:00 |
| 363 | ['rt', 'hasar | vaexdanny | 2023-04-10 01:57:13+00:00 |
| 218 | ['�', 'enha | torksmith | 2023-04-10 02:03:03+00:00 |
| 332 | ['rt', '0xgau | WeASeL_A | 2023-04-10 01:58:29+00:00 |
| 412 | ['rt', 'abhis | parasher_r | 2023-04-10 01:55:14+00:00 |
| 977 | ['jdonthero | Rgr_Tht_ | 2023-04-10 01:31:19+00:00 |
| 120 | ['rt', 'uberf | m_dsemw | 2023-04-10 02:06:41+00:00 |
| 964 | ['2', 'type', | Marta_Lya | 2023-04-10 01:31:47+00:00 |
| 535 | ['rt', 'workl | OffOfOnHe | 2023-04-10 01:50:04+00:00 |
| 697 | ['rt', 'therui | parvez1 | 2023-04-10 01:42:58+00:00 |
| 686 | ['india', 'pl | DeepakNe | 2023-04-10 01:43:25+00:00 |
| 408 | ['rt', 'briani | justinthemi | 2023-04-10 01:55:18+00:00 |
| 55 | ['rt', 'frkad | venikunche | 2023-04-10 02:09:12+00:00 |
| 155 | ['7/', 'strat | Debabrata | 2023-04-10 02:05:18+00:00 |
| 953 | ['rt', 'pape | Dharma09 | 2023-04-10 01:32:10+00:00 |
| 215 | ['rt', 'nntale | rohitdhawa | 2023-04-10 02:03:04+00:00 |
| 240 | ['ericlewis' | GodlyIgno | 2023-04-10 02:02:20+00:00 |
| 882 | ['rt', 'thesh | skmani380 | 2023-04-10 01:35:01+00:00 |
| 207 | ['rt', 'miran | Squirrel11 | 2023-04-10 02:03:17+00:00 |

EDA:
Not yet.

We are planning to use decision trees or random forests to analyze the dataset in the future.

## Coding

Data Mining & Preprocessing: Yuehan Qin, Ming Tang

Modeling: Jianhui Ding, Bofei Wang

Future work:

Model evaluation and improvement: Everyone

Code cleanup and documentation: Everyone