

## Document 2. Twitter Scraper using Selenium

### Aim

To identify gender bias in science, technology, engineering and mathematics (STEM) in language data and prototype a method to identify the source of this bias, in order to address it.

### Rationale

There is a gender imbalance in STEM careers (Robnett, 2016). In 2016, women were particularly underrepresented in IT and engineering, comprising 28% and 12.4% of the workforce, respectively (Australian Government, 2019). Because gender disparities are rooted in cultural norms, values and discourse (Robnett, 2016), it is expected that this imbalance should appear as bias in language data (Sun et al., 2019; Zhao et al., 2018). This bias may even be amplified in any predictions made based on language data (Sun et al. 2019).

Natural Language Processing (NLP) is the application of data science and analytics to human language (Chowdhury, 2003). Language comprises some of the richest and most complex data available, and can provide nuanced insights to virtually any topic, in terms of human thoughts and reactions to different phenomena (Chowdhury, 2003; Hirschberg and Manning, 2015). Unsurprisingly, the rise of social media has provided a bounty of data for NLP applications. In 2020, 500 million Tweets were published daily on social media site Twitter, providing an extensive range of language data on a multitude of different topics (Sayce, 2020). Social media platforms, such as Twitter, behave as a largely uncensored, global-scale historical record from the perspective of everyday members of society (Neethu, 2013), not just the powerful and privileged, as has been the case in the past (although, importantly, and in line with the topic of this report, Tweets written by people of colour are more likely to be flagged as offensive and removed by machine learning algorithms (Wei, 2020)).

Because this investigation seeks to research implicit gender bias in STEM-related topics, and because gender bias is a social issue, it is sensible to look for data where people produce a high volume of text. Because Twitter can be treated as a repository of people's opinions and attitudes, and therefore their personal biases, it naturally harbours the biases of society in general, making Twitter an ideal source of data for sentiment analysis (Neethu, 2013).

Scraping data from Twitter, using specific search terms, allows targeted collection of information and ensures that the data collected are relevant to this investigation. For these reasons, Twitter is the sole website from which data were extracted.

```
In [3]: # Import packages
import pandas as pd
import re
import time
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.common.exceptions import NoSuchElementException
from selenium.common.exceptions import StaleElementReferenceException
from webdriver_manager.chrome import ChromeDriverManager
import csv
import re
import string
import nltk
import nltk.stem
import nltk.stem.WordNetLemmatizer
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
from matplotlib.pyplot import ticks
import matplotlib.ticker as mtick
from matplotlib.ticker import StrMethodFormatter
```

### Methodology

This report will analyse language data bias in STEM, with respect to gender, using Python version 3.9.0. The methodology is as follows:

- Scrape tweets about STEM, relating to people.
- Clean the data so it can be used in NLP tasks.
- Investigate the number of STEM-related tweets containing female pronouns vs male pronouns.
- Report findings from a preliminary exploration of these data.
- Identify the source of bias using a combination of supervised classification and sentiment analysis.

Although Twitter has an API which can be used for academic purposes, in line with the task requirements Selenium was used to simulate a search within a browser and data were collected using `xpath()`. The search term was `(science OR scientist OR technology OR technician OR engineering OR engineer OR math OR mathematics OR mathematics OR math OR her OR he OR him) -filter: retweets`.

### Webscraper

The data from Twitter were complicated to extract because the site is complex and the content is continuously updating. Due to the perpetual scrolling nature of the site, it was decided that the best way to scrape Tweet data (other than using the API) would be using the Selenium package. Selenium works by automating web browser activity, replicating functions such as opening a browser, clicking on links and scrolling pages (Silman, 2019). The package requires installation of a driver to interface with a browser (Chrome was used in this case).

Selenium was used to open, log in, navigate and enter a search term to scrape Tweet data while scrolling downwards, allowing new Tweets to appear. The `xpath()` function was used to select specific data from each Tweet. Username, handle, Tweet and the number of replies, retweets, likes and the post date were retrieved. Figure 1 shows the webscraper in progress.

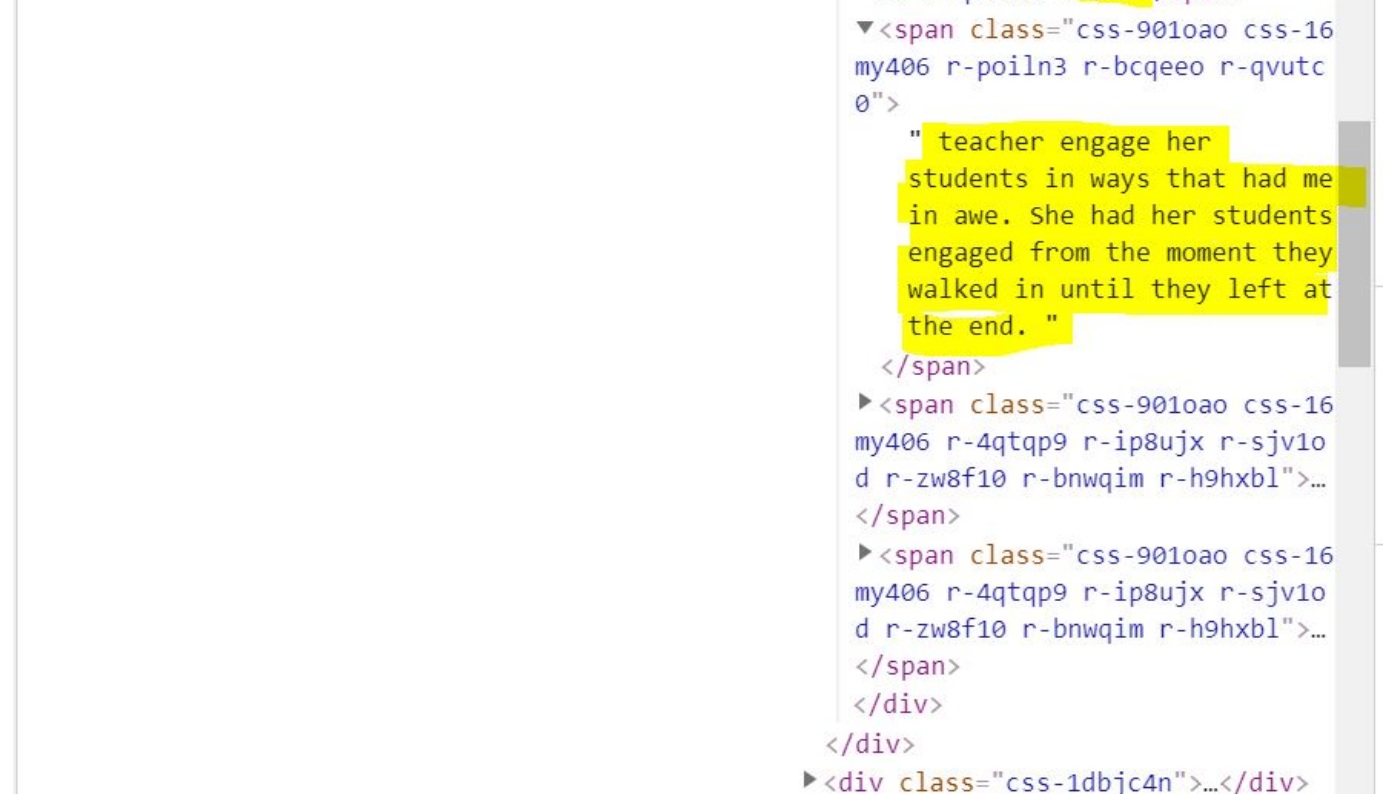


Figure 1: The webscraper in action

After identifying issues with the running of the webscraper, modifications were made to the code to catch two types of exception: `NoSuchElementException`, which can occur when the scraper attempts to access an element before the page has fully loaded, and `StaleElementReferenceException`, which occurs when some element is temporarily inaccessible. Elements were identified and extracted using the HTML inspect navigation panel corresponding to the page in the Chrome browser (Figure 2):



Figure 2: Tweet within inspect panel

```
In [2]: # Function to scrape Twitter using Selenium
def get_tweet_data(card):
    #Extract tweet data
    username = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    try:
        user_name = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #Twitter handle
    try:
        handle = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #tweet text
    try:
        comment = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
        responding = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
        text = comment+responding
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #reply count
    try:
        comment = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #retweet count
    try:
        retweet = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #likes
    try:
        like = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    #post date
    try:
        date = card.find_element_by_xpath('//*[@name=session[username_or_email]]').text
    except NoSuchElementException:
        return
    except StaleElementReferenceException:
        return
    tweet = (username, handle, text, comment, retweet, like, date)
    return tweet
driver = webdriver.Chrome()
driver.get('https://www.twitter.com/login')
```

```
In [3]: #navigate to twitter and login
username = driver.find_element_by_xpath('//*[@name=session[username_or_email]]').text
password = driver.find_element_by_xpath('//*[@name=session[password]]').text
password.send_keys(password)
password.send_keys(Keys.RETURN)
.....
```

```
In [5]: #find search input and search for term
#navigate to 'explore' tab
driver.find_element_by_xpath('//*[@name=session[username_or_email]]').click()
search_input = driver.find_element_by_xpath('//*[@name=session[username_or_email]]').text
search_input.send_keys('science OR technology OR engineering OR math OR mathematics OR science OR technology OR engineering OR mathematician (she OR her OR he OR him) -filter:retweets')
search_input.send_keys(Keys.RETURN)
#navigate to 'latest' tab
driver.find_element_by_link_text('Latest').click()
# get all tweets on the page
data = []
tweet_ids = set()
last_position = driver.execute_script("return window.pageYOffset;")
scrolling = True
while scrolling:
    page_cards = driver.find_elements_by_xpath('//*[@name=session[username_or_email]]')
    for card in page_cards[:15]:
        tweet = get_tweet_data(card)
        if tweet:
            tweet_id = ''.join(tweet)
            if tweet_id not in tweet_ids:
                tweet_ids.add(tweet_id)
                data.append(tweet)
    scroll_attempt = 0
    while True:
        driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
        sleep(1)
        curr_position = driver.execute_script("return window.pageYOffset;")
        if last_position == curr_position:
            scroll_attempt += 1
        #end of scroll region
        if scroll_attempt == 3:
            scrolling = False
            break
        else:
            sleep(2) # attempt to scroll again
            last_position = curr_position
            break
#save tweet data
with open('stem_tweets_2.csv', 'w', newline='') as f:
    header = ['Username', 'Handle', 'Text', 'Comments', 'Retweets', 'Likes', 'Date']
    writer = csv.writer(f)
    writer.writerow(header)
    writer.writerows(data)
```

### Viewing the scraped data

Once the scraper had collected all the data, the information was saved to a .csv file. Table 1 shows the resulting dataset comprised over 34,000 records of STEM-related tweets from April 14, 2021 (the day of collection).

```
In [3]: # Import data and view dataframe head
tweets_df = pd.read_csv('stem_tweets.csv', dtype=object)
tweets_df = tweets_df.fillna(0)
tweets_df.head(10)
```

	Username	Handle	Text	Comments	Retweets	Likes	Date
0	Angi	@AngiMayssa	Blackwell is great. As a lawyer, you have to b...		0	0	2021-04-14T20:32:29.000Z
1	Safer Luis (LUV)	@lucascab69	math dumb as if i never then ymf don't she la...		0	0	2021-04-14T20:32:24.000Z
2	Jared Halpern	@JaredHalpern	Wow, he was a hell of an engineer - Great brain...		0	0	2021-04-14T20:32:23.000Z
3	Bob Brigham	@BobBrigham	Matt Kelly is so bad he thinks coronavirus is...		0	0	2021-04-14T20:32:10.000Z
4	AlaskanTzar	@AlaskanTzar	Replying to @BobbBrigham was? i'm absolutely bl...		0	0	2021-04-14T20:31:55.000Z
5	worst vovne (enka)	@vovne_beshe	that math test i just took was by absolute bl...		0	0	2021-04-14T20:31:55.000Z
6	Thomas Howard Riley MAD SCIENTIST OF THE FANTASY	@Gomthepoxyx	Replying to @vovne_beshe is so funny...		0	0	2021-04-14T20:31:43.000Z
7	m d ryan	@haskinsryan526	Replying to @vovne_beshe and @BobbBrigham...		0	0	2021-04-14T20:31:39.000Z
8	Stephen	@Steve_PoChange	Replying to @BobbBrigham @StephenKewen and...		0	0	2021-04-14T20:31:33.000Z
9	catady	@suchmeekak	Replying to @vovne_beshe and @BobbBrigham...		0	0	2021-04-14T20:31:18.000Z

### Data cleaning

There were several necessary steps involved in preparing the Tweet data for analysis using NLP methods. First, user handles were removed using a lambda function targeting the @ symbol before user handles. After this, the remaining punctuation, special characters were removed. Furthermore, a new column was generated with values 'M' or 'F', corresponding to male or female, depending on which pronouns were present in the original Tweet, and a final 'Topic' column containing the relevant STEM field.

The resulting dataframe contained 25,365 rows, with four columns containing the User Handle, the text of each Tweet, and the subject gender (Table 2). The dataframe was saved as a .csv file for later use.

```
In [10]: # Remove handles
tweets_df['Text'] = tweets_df['Text'].apply(lambda x: re.sub('@(w|u|)', '', str(x)))

# Remove punctuation
tweets_df['Text'] = tweets_df['Text'].str.replace('!', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('.', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(',', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(';', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(':', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('\"', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('\'', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('`', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('~', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('_', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('-', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(' ', '')
tweets_df['Text'] = tweets_df['Text'].str.replace('(', '')
tweets_df['Text'] = tweets_df['Text'].str.replace(')',
```