IS567 - Text Mining
Assignment 1

Task 1: Load the dataset

Task 1: Loading the dataset

In [3]:
```
chatgpt_tweets = pd.read_csv("chatgpt.csv")
chatgpt_tweets.head()
```

Out[3]:

	Datetime	Tweet Id	Text	Username	Permalink	User
0	2023-01-22 13:44:34+00:00	1617156270871699456	ChatGPTで遊ぶの忘れてた！！\n書類作るコード書いてみてほしいのと、\nどこまで思考整...	mochico0123	https://twitter.com/mochico0123/status/1617156...	https://twitter.com/mochico0123
1	2023-01-22 13:44:39+00:00	1617156291046133761	@Alexandrovnalng Prohibition of ChatGPT has be...	Caput_LupinumSG	https://twitter.com/Caput_LupinumSG/status/161...	https://twitter.com/Caput_LupinumSG
2	2023-01-22 13:44:44+00:00	1617156308926349312	Schaut Euch an, was @fobizz @DianaKnodel alles...	ciffi	https://twitter.com/ciffi/status/1617156308926...	https://twitter.com/ciffi
3	2023-01-22 13:44:49+00:00	1617156332297256961	Bow down to chatGPT 👊..... https://t.co/ENTSzi...	Vishwasrisiri	https://twitter.com/Vishwasrisiri/status/16171...	https://twitter.com/Vishwasrisiri
4	2023-01-22 13:44:52+00:00	1617156345064570880	Profilinde vatan, Türkiye falan yazan bireyler...	0xGenetikciniz	https://twitter.com/0xGenetikciniz/status/1617...	https://twitter.com/0xGenetikciniz

The column names in the dataset

In [4]:
```
chatgpt_tweets.columns
```

Out[4]:
```
Index(['Datetime', 'Tweet Id', 'Text', 'Username', 'Permalink', 'User',
       'Outlinks', 'CountLinks', 'ReplyCount', 'RetweetCount', 'LikeCount',
       'QuoteCount', 'ConversationId', 'Language', 'Source', 'Media',
       'QuotedTweet', 'MentionedUsers', 'hashtag', 'hastag_counts'],
      dtype='object')
```

Number of instances in the dataset

In [5]:
```
len(chatgpt_tweets)
```

Out[5]:
```
50001
```

As we see the last output, we see that after loading the dataset we have **50001** instances of tweets.

Task 2: Remove all non-English Tweets

Task 2: Remove all non-English tweets

```
In [7]: only_english_tweets = chatgpt_tweets[chatgpt_tweets['Language'] == 'en']
```

```
In [8]: only_english_tweets.head()
```

Out[8]:

	Datetime	Tweet Id	Text	Username	Permalink	User
1	2023-01-22 13:44:39+00:00	1617156291046133761	@Alexandrovnalng Prohibition of ChatGPT has be...	Caput_LupinumSG	https://twitter.com/Caput_LupinumSG/status/161...	https://twitter.com/Caput_LupinumSG
3	2023-01-22 13:44:49+00:00	1617156332297256961	Bow down to chatGPT 👊..... https://t.co/ENTSzi...	Vishwasrisiri	https://twitter.com/Vishwasrisiri/status/16171...	https://twitter.com/Vishwasrisiri
6	2023-01-22 13:45:03+00:00	1617156389217894400	ChatGPT runs 10K Nvidia training GPUs with pot...	FierceElectron	https://twitter.com/FierceElectron/status/1617...	https://twitter.com/FierceElectron
7	2023-01-22 13:45:04+00:00	1617156393898745858	@SWENGDAD There is repetitive work in every jo...	xlr8harder	https://twitter.com/xlr8harder/status/16171563...	https://twitter.com/xlr8harder
8	2023-01-22 13:45:06+00:00	1617156404137295878	I created a fictional jewelry brand using Chat...	Kosuke_dazo	https://twitter.com/Kosuke_dazo/status/1617156...	https://twitter.com/Kosuke_dazo

```
In [7]: only_english_tweets['Language'].unique()
```

```
Out[7]: array(['en'], dtype=object)
```

Number of instances after removing non-English tweets

```
In [8]: len(only_english_tweets)
```

```
Out[8]: 32076
```

As we see the last output, after removal of non-English tweets we have **32076** tweets remaining.

Task 3: Perform basic transformations

Convert to lowercase

```
only_english_tweets[['Text']]
```

	Text
1	@alexandrovnaing prohibition of chatgpt has been added to the honor code of my daughter's school
3	bow down to chatgpt 👈..... https://t.co/entszi2aq9
6	chatgpt runs 10k nvidia training gpus with potential for thousands more https://t.co/uhq62t0uw4
7	@swengdad there is repetitive work in every job, there are lots of small tasks that can leverage chatgpt or copilot and keep you working at a higher level of abstraction.
8	i created a fictional jewelry brand using chatgpt and midjourney.\n\nhttps://t.co/gtwdnz0lam\n\n#chatgpt #midjourney https://t.co/n5hlelzpty
...	...
49991	i joined the @aipadtech x @moonsalecom exclusive giveaway for a chance to win being a part of the #aipad launch! get ready for a launch like no other!\n\ntg: https://t.co/5if0uosveq\n\n btceth #ai #artificialintelligence #crypto #chatgpt #future https://t.co/kp71xazn14
49992	@iamjohnoliver i think chatgpt is taking the piss. #lastweektonight https://t.co/tvlmpntlbw
49993	digital marketers adopt new skills. artificial intelligence is completely going to capture your jobs. in a few months there will be no work for you in industry #chatgpt \n\nseo\nsmm\ncontent creators\ngraphic designer\nadvertising\nbpo/kpo\n\nabove 90% to 95% chances to drop out.
49995	remember when @twitter was down a lot in the early years cause too many people were using it. @openai is having the same problem with #chatgpt now. a shift is here.
49999	portland shop uses chatgpt to tell family stories on a startup budget https://t.co/rzgvr6ytoc

32076 rows × 1 columns

Separate hashtags (#) and account name (@) from the actual tweet text

In [51]:
```
only_english_tweets.loc[:, ('Hashtags')] = only_english_tweets.loc[:, ('Text')].apply(separate_hashtags)
only_english_tweets.loc[:, ('Accounts')] = only_english_tweets.loc[:, ('Text')].apply(separate_account_name)
```

In [37]:
```
only_english_tweets[['Text', 'Hashtags', 'Accounts']]
```

Out[37]:

	Text	Hashtags	Accounts
1	@alexandrovnaing prohibition of chatgpt has be...	[]	[alexandrovnaing]
3	bow down to chatgpt 👈..... https://t.co/entszi...	[]	[]
6	chatgpt runs 10k nvidia training gpus with pot...	[]	[]
7	@swengdad there is repetitive work in every jo...	[]	[swengdad]
8	i created a fictional jewelry brand using chat...	[chatgpt, midjourney]	[]
...	...	...	...
49991	i joined the @aipadtech x @moonsalecom exclusi...	[aipad, ai, artificialintelligence, crypto, ch...	[aipadtech, moonsalecom]
49992	@iamjohnoliver i think chatgpt is taking the p...	[lastweektonight]	[iamjohnoliver]
49993	digital marketers adopt new skills. artificial...	[chatgpt]	[]
49995	remember when @twitter was down a lot in the e...	[chatgpt]	[twitter, openai]
49999	portland shop uses chatgpt to tell family stor...	[]	[]

32076 rows × 3 columns

Remove non-alphanumeric characters that appear in CleanedTweet

```
In [64]: only_english_tweets.loc[:, ('CleanedTweet')] = only_english_tweets.loc[:, ('Text')].apply(clean_tweet)
```

```
In [61]: only_english_tweets[['CleanedTweet']]
```

Out[61]:

	CleanedTweet
1	prohibition of chatgpt has been added to the honor code of my daughter s school
3	bow down to chatgpt https t co entszi2aq9
6	chatgpt runs 10k nvidia training gpus with potential for thousands more https t co uhq62t0uw4
7	there is repetitive work in every job there are lots of small tasks that can leverage chatgpt or copilot and keep you working at a higher level of abstraction
8	i created a fictional jewelry brand using chatgpt and midjourney https t co gtwdnz0lam https t co n5hlelzpty
...	...
49991	i joined the x exclusive giveaway for a chance to win being a part of the launch get ready for a launch like no other tg https t co 5if0uosveq btc eth https t co kp71xazn14
49992	i think chatgpt is taking the piss https t co tvlmpntlbw
49993	digital marketers adopt new skills artificial intelligence is completely going to capture your jobs in a few months there will be no work for you in industry seo smm content creators graphic designer advertising bpo kpo above 90 to 95 chances to drop out
49995	remember when was down a lot in the early years cause too many people were using it is having the same problem with now a shift is here
49999	portland shop uses chatgpt to tell family stories on a startup budget https t co rzgvr6ytoc

32076 rows × 1 columns

3 examples of Tweets before and after Text processing

3 tweets before and after text processing

```
only_english_tweets.loc[3:3, ['Text', 'CleanedTweet', 'Hashtags', 'Accounts']]
```

	Text	CleanedTweet	Hashtags	Accounts
3	bow down to chatgpt 🫳..... https://t.co/entszi2aq9	bow down to chatgpt https t co entszi2aq9	[]	[]

```
only_english_tweets.loc[42360:42360, ['Text', 'CleanedTweet', 'Hashtags', 'Accounts']]
```

	Text	CleanedTweet	Hashtags	Accounts
42360	@kathleenskeals now i'm going to have to use chatgpt to create awesome tweets! https://t.co/v02gubsyib	now i m going to have to use chatgpt to create awesome tweets https t co v02gubsyib	[]	[kathleenskeals]

```
only_english_tweets.loc[23000:23000, ['Text', 'CleanedTweet', 'Hashtags', 'Accounts']]
```

	Text	CleanedTweet	Hashtags	Accounts
23000	#chatgpt #ai "yall ain't ready" https://t.co/0ycgqkobrp	yall ain t ready https t co 0ycgqkobrp	[chatgpt, ai]	[]

Task 4: Tokenization (two approaches)

1) Using Python NLTK:

```
only_english_tweets[['TokenizedTweetNLTK']]
```

	TokenizedTweetNLTK
1	[prohibition, of, chatgpt, has, been, added, to, the, honor, code, of, my, daughter, s, school]
3	[bow, down, to, chatgpt, https, t, co, entszi2aq9]
6	[chatgpt, runs, 10k, nvidia, training, gpus, with, potential, for, thousands, more, https, t, co, uhq62t0uw4]
7	[there, is, repetitive, work, in, every, job, there, are, lots, of, small, tasks, that, can, leverage, chatgpt, or, copilot, and, keep, you, working, at, a, higher, level, of, abstraction]
8	[i, created, a, fictional, jewelry, brand, using, chatgpt, and, midjourney, https, t, co, gtwdnz0lam, https, t, co, n5hlelzpty]
...	...
49991	[i, joined, the, x, exclusive, giveaway, for, a, chance, to, win, being, a, part, of, the, launch, get, ready, for, a, launch, like, no, other, tg, https, t, co, 5if0uosveq, btc, eth, https, t, co, kp71xazn14]
49992	[i, think, chatgpt, is, taking, the, piss, https, t, co, tvlmpntlbw]
49993	[digital, marketers, adopt, new, skills, artificial, intelligence, is, completely, going, to, capture, your, jobs, in, a, few, months, there, will, be, no, work, for, you, in, industry, seo, smm, content, creators, graphic, designer, advertising, bpo, kpo, above, 90, to, 95, chances, to, drop, out]
49995	[remember, when, was, down, a, lot, in, the, early, years, cause, too, many, people, were, using, it, is, having, the, same, problem, with, now, a, shift, is, here]
49999	[portland, shop, uses, chatgpt, to, tell, family, stories, on, a, startup, budget, https, t, co, rzgvr6ytoc]

32076 rows × 1 columns

2) Using SpaCy:

```
only_english_tweets[['TokenizedTweetSpacy']]
```

	TokenizedTweetSpacy
1	[prohibition, of, chatgpt, has, been, added, to, the, honor, code, of, my, daughter, s, school]
3	[bow, down, to, chatgpt, , https, , t, co, entszi2aq9]
6	[chatgpt, runs, 10k, nvidia, training, gpus, with, potential, for, thousands, more, https, , t, co, uhq62t0uw4]
7	[there, is, repetitive, work, in, every, job, , there, are, lots, of, small, tasks, that, can, leverage, chatgpt, or, copilot, and, keep, you, working, at, a, higher, level, of, abstraction]
8	[i, created, a, fictional, jewelry, brand, using, chatgpt, and, midjourney, , https, , t, co, gtwdnz0lam, https, , t, co, n5hlelzpty]
...	...
49991	[i, joined, the, x, exclusive, giveaway, for, a, chance, to, win, being, a, part, of, the, launch, , get, ready, for, a, launch, like, no, other, , tg, , https, , t, co, 5if0uosveq, , btc, , eth, https, , t, co, kp71xazn14]
49992	[i, think, chatgpt, is, taking, the, piss, , https, , t, co, tvlmpntlbw]
49993	[digital, marketers, adopt, new, skills, , artificial, intelligence, is, completely, going, to, capture, your, jobs, , in, a, few, months, there, will, be, no, work, for, you, in, industry, seo, smm, content, creators, graphic, designer, advertising, bpo, kpo, above, 90, , to, 95, , chances, to, drop, out]
49995	[remember, when, was, down, a, lot, in, the, early, years, cause, too, many, people, were, using, it, , is, having, the, same, problem, with, now, , a, shift, is, here]
49999	[portland, shop, uses, chatgpt, to, tell, family, stories, on, a, startup, budget, https, , t, co, rzgvr6ytoc]

32076 rows × 1 columns

Example:

```
only_english_tweets.loc[23000, ['Text', 'CleanedTweet', 'TokenizedTweetNLTK', 'TokenizedTweetSpacy']]
```

```
Text                    #chatgpt #ai  "yall ain't ready" https://t.co/0ycgqkobrp
CleanedTweet                           yall ain t ready  https   t co 0ycgqkobrp
TokenizedTweetNLTK                 [yall, ain, t, ready, https, t, co, 0ycgqkobrp]
TokenizedTweetSpacy    [ , y, all, ain, t, ready,  , https,   , t, co, 0ycgqkobrp]
Name: 23000, dtype: object
```

Observations:

A) Differences:

NLTK Tokenizer	SpaCy Tokenizer
- The NLTK tokenizer makes use of rule-based tokenization.	- The spaCy tokenizer uses statistics for boundary determination.
- It treats the white spaces (spaces, tabs or new lines) as delimiters and splits when it encounters one. NLTK discards the extra whitespace characters and multiple consecutive whitespaces will not be retained in the NLTK tokenised output.	- SpaCy handles whitespaces little differently and the whitespace characters are retained. Multiple whitespaces, if encountered are not touched and the structure remains the same.
- NLTK does not handle the contracted words well like I'm or ain't	- SpaCy can handle the contractions well and splits them separately.
- NLTK tokenization takes more time and is slower	- SpaCy tokenization is faster

B) Similarities:

- Both Python NLTK tokenizer and SpaCy tokenizer are toolkits designed to handle word tokenization for text processing.
- Both can handle a large variety of text and support multiple languages.
- They are easy to use as both have support for simple APIs and functions that can be leveraged to tokenise text in Python.

Based on the above observations, I prefer the SpaCy tokenization and will use that for the rest of the processing steps as:

- It does not discard the multiple consecutive whitespace characters and this may be useful in retaining the original text structure.
- SpaCy has better text tokenization algorithm as it can split contractions as well
- SpaCy tokenization is faster and has a better performance

Task 5: Remove stopwords

We store the tokens in the TokenizedTweet after using SpaCy tokenizer

```
: from spacy.lang.en.stop_words import STOP_WORDS

: STOP_WORDS
```

```
'would',
'yet',
'you',
'your',
'yours',
'yourself',
'yourselves',
''d',
''ll',
''m',
''re',
''s',
''ve',
'’d',
'’ll',
'’m',
'’re',
'’s',
'’ve'}
```

```
only_english_tweets[['StopwordRemovedTweet']]
```

	StopwordRemovedTweet
1	[prohibition, chatgpt, added, honor, code, daughter, s, school]
3	[bow, chatgpt, , https, , t, co, entszi2aq9]
6	[chatgpt, runs, 10k, nvidia, training, gpus, potential, thousands, https, , t, co, uhq62t0uw4]
7	[repetitive, work, job, , lots, small, tasks, leverage, chatgpt, copilot, working, higher, level, abstraction]
8	[created, fictional, jewelry, brand, chatgpt, midjourney, , https, , t, co, gtwdnz0lam, https, , t, co, n5hlelzpty]
...	...
49991	[joined, x, exclusive, giveaway, chance, win, launch, , ready, launch, like, , tg, , https, , t, co, 5if0uosveq, , btc, , eth, https, , t, co, kp71xazn14]
49992	[think, chatgpt, taking, piss, , https, , t, co, tvlmpntlbw]
49993	[digital, marketers, adopt, new, skills, , artificial, intelligence, completely, going, capture, jobs, , months, work, industry, seo, smm, content, creators, graphic, designer, advertising, bpo, kpo, 90, , 95, , chances, drop]
49995	[remember, lot, early, years, cause, people, , having, problem, , shift]
49999	[portland, shop, uses, chatgpt, tell, family, stories, startup, budget, https, , t, co, rzgvr6ytoc]

32076 rows × 1 columns

Example:

```
only_english_tweets.loc[3:3, ['CleanedTweet', 'TokenizedTweet', 'StopwordRemovedTweet']]
```

	CleanedTweet	TokenizedTweet	StopwordRemovedTweet
3	bow down to chatgpt https t co entszi2aq9	[bow, down, to, chatgpt, , https, , t, co, entszi2aq9]	[bow, chatgpt, , https, , t, co, entszi2aq9]

Task 6: Stemming

1) Using Python NLTK Porter stemmer:

```
only_english_tweets[['StemmedTweet']]
```

	StemmedTweet
1	[prohibit, of, chatgpt, ha, been, ad, to, the, honor, code, of, my, daughter, s, school]
3	[bow, down, to, chatgpt, , http, , t, co, entszi2aq9]
6	[chatgpt, run, 10k, nvidia, train, gpu, with, potenti, for, thousand, more, http, , t, co, uhq62t0uw4]
7	[there, is, repetit, work, in, everi, job, , there, are, lot, of, small, task, that, can, leverag, chatgpt, or, copilot, and, keep, you, work, at, a, higher, level, of, abstract]
8	[i, creat, a, fiction, jewelri, brand, use, chatgpt, and, midjourney, , http, , t, co, gtwdnz0lam, http, , t, co, n5hlelzpti]
...	...
49991	[i, join, the, x, exclus, giveaway, for, a, chanc, to, win, be, a, part, of, the, launch, , get, readi, for, a, launch, like, no, other, , tg, , http, , t, co, 5if0uosveq, , btc, , eth, http, , t, co, kp71xazn14]
49992	[i, think, chatgpt, is, take, the, piss, , http, , t, co, tvlmpntlbw]
49993	[digit, market, adopt, new, skill, , artifici, intellig, is, complet, go, to, captur, your, job, , in, a, few, month, there, will, be, no, work, for, you, in, industri, seo, smm, content, creator, graphic, design, advertis, bpo, kpo, abov, 90, , to, 95, , chanc, to, drop, out]
49995	[rememb, when, wa, down, a, lot, in, the, earli, year, caus, too, mani, peopl, were, use, it, , is, have, the, same, problem, with, now, , a, shift, is, here]
49999	[portland, shop, use, chatgpt, to, tell, famili, stori, on, a, startup, budget, http, , t, co, rzgvr6ytoc]

32076 rows × 1 columns

2) Using Python NLTK Lancaster stemmer:

```
only_english_tweets[['StemmedTweetLancaster']]
```

	StemmedTweetLancaster
1	[prohibit, of, chatgpt, has, been, ad, to, the, hon, cod, of, my, daught, s, school]
3	[bow, down, to, chatgpt, , https, , t, co, entszi2aq9]
6	[chatgpt, run, 10k, nvid, train, gpu, with, pot, for, thousand, mor, https, , t, co, uhq62t0uw4]
7	[ther, is, repetit, work, in, every, job, , ther, ar, lot, of, smal, task, that, can, lev, chatgpt, or, copilot, and, keep, you, work, at, a, high, level, of, abstract]
8	[i, cre, a, fict, jewelry, brand, us, chatgpt, and, midjourney, , https, , t, co, gtwdnz0lam, https, , t, co, n5hlelzpty]
...	...
49991	[i, join, the, x, exclud, giveaway, for, a, chant, to, win, being, a, part, of, the, launch, , get, ready, for, a, launch, lik, no, oth, , tg, , https, , t, co, 5if0uosveq, , btc, , eth, https, , t, co, kp71xazn14]
49992	[i, think, chatgpt, is, tak, the, piss, , https, , t, co, tvlmpntlbw]
49993	[digit, market, adopt, new, skil, , art, intellig, is, complet, going, to, capt, yo, job, , in, a, few, month, ther, wil, be, no, work, for, you, in, industry, seo, smm, cont, cre, graph, design, advert, bpo, kpo, abov, 90, , to, 95, , chant, to, drop, out]
49995	[rememb, when, was, down, a, lot, in, the, ear, year, caus, too, many, peopl, wer, us, it, , is, hav, the, sam, problem, with, now, , a, shift, is, her]
49999	[portland, shop, us, chatgpt, to, tel, famy, story, on, a, startup, budget, https, , t, co, rzgvr6ytoc]

32076 rows × 1 columns

Example:

3 examples for StemmedTweet

```
only_english_tweets.loc[3:3, ['CleanedTweet', 'TokenizedTweet', 'StemmedTweet']]
```

	CleanedTweet	TokenizedTweet	StemmedTweet
3	bow down to chatgpt https t co entszi2aq9	[bow, down, to, chatgpt, , https, , t, co, entszi2aq9]	[bow, down, to, chatgpt, , http, , t, co, entszi2aq9]

```
only_english_tweets.loc[23000:23000, ['CleanedTweet', 'TokenizedTweet', 'StemmedTweet']]
```

	CleanedTweet	TokenizedTweet	StemmedTweet
23000	yall ain t ready https t co 0ycgqkobrp	[, y, all, ain, t, ready, , https, , t, co, 0ycgqkobrp]	[, y, all, ain, t, readi, , http, , t, co, 0ycgqkobrp]

```
only_english_tweets.loc[42360:42360, ['CleanedTweet', 'TokenizedTweet', 'StemmedTweet']]
```

	CleanedTweet	TokenizedTweet	StemmedTweet
42360	now i m going to have to use chatgpt to create awesome tweets https t co v02gubsyib	[now, i, m, going, to, have, to, use, chatgpt, to, create, awesome, tweets, , https, , t, co, v02gubsyib]	[now, i, m, go, to, have, to, use, chatgpt, to, creat, awesom, tweet, , http, , t, co, v02gubsyib]

Observations:

A) Differences:

Porter Stemmer	Lancaster Stemmer
- The Porter stemmer uses less rules and is not aggressive in stemming.	- The Lancaster stemmer uses an extensive set of rules and is generally aggressive in stemming.
- Since it does not apply very aggressive stemming, the results are more accurate when it comes to keeping the original meaning intact for the stemmed words.	- Due to aggressive stemming, it may sometimes lose the original word and give a root that may not fit in the text processing context.
- Porter stemming is much faster but in the test example, I found it to be slower than Lancaster stemming.	- Lancaster stemming is slower but in the test example, I found it to be faster than Porter stemming.

B) Similarities:

- Both Python NLTK Porter and Lancaster stemmers are toolkits designed to simplify words to their common roots for text processing.
- Both stemmers use rule-based algorithms for stemming.
- Both stemmers are designed for stemming English language words and unsuited for other languages.
- They are easy to use as both have support for simple APIs and functions that can be leveraged to stem text in Python.

Narrative Questions:

Observations on the dataset:

- When we find the descriptive statistics of the numerical columns, I observed that the median value for columns like ReplyCount, LikeCount, RetweetCount, QuoteCount and hastag_counts was 0, indicating that many tweets were not replied to, liked, retweeted or quoted.
- There are no duplicate tweets.
- Maximum document length for tweets is 826, minimum length is 7 and on an average there are 149 words per document.
- From the word cloud after removing stop words and non-alphanumeric characters, the most common words are https, chatgpt, t, co as obvious because the tweets were links and related to chatgpt.

Minimum, maximum and average document length

```
maximum_doc_length = only_english_tweets['Text'].str.len().max()
maximum_doc_length
```

```
826
```

```
minimum_doc_length = only_english_tweets['Text'].str.len().min()
minimum_doc_length
```

```
7
```

```
average_doc_length = only_english_tweets['Text'].str.len().mean()
average_doc_length
```

```
148.71480234443197
```

Observations on the data preprocessing steps:

- The general pipeline to process text begins with sentence segmentation. This was not done in this task.
- Some columns like date or Source do not have any value and do not add any special meaning when converting to features. Such columns should be dropped to reduce the dataset size. But they were not dropped in this task.
- Columns that may have null values and are important to the final use-case should be handled to substitute the null values with placeholder values.

Other preprocessing steps to aid in text cleaning:

- Sentence Segmentation
- Lemmatization
- Handling of numbers
- Handling of punctuation marks
- Multiword expressions
- Removing HTML tags
- Handling contractions
- Handling rare or domain specific words

Bonus Task:

Sentence Segmentation:

```
only_english_tweets.loc[:, ('SentenceSegmentedTweet')] = only_english_tweets.loc[:, ('Text')].apply(sentence_segmentat
```

```
only_english_tweets[['Text', 'SentenceSegmentedTweet']]
```

	Text	SentenceSegmentedTweet
1	@alexandrovnaing prohibition of chatgpt has been added to the honor code of my daughter's school	[@alexandrovnaing prohibition of chatgpt has been added to the honor code of my daughter's school]
3	bow down to chatgpt 🫡..... https://t.co/entszi2aq9	[bow down to chatgpt 🫡..... https://t.co/entszi2aq9]
6	chatgpt runs 10k nvidia training gpus with potential for thousands more https://t.co/uhq62t0uw4	[chatgpt runs 10k nvidia training gpus with potential for thousands more https://t.co/uhq62t0uw4]
7	@swengdad there is repetitive work in every job, there are lots of small tasks that can leverage chatgpt or copilot and keep you working at a higher level of abstraction.	[@swengdad there is repetitive work in every job, there are lots of small tasks that can leverage chatgpt or copilot and keep you working at a higher level of abstraction.]
8	i created a fictional jewelry brand using chatgpt and midjourney.\n\nhttps://t.co/gtwdnz0lam\n\n#chatgpt #midjourney https://t.co/n5hlelzpty	[i created a fictional jewelry brand using chatgpt and midjourney., https://t.co/gtwdnz0lam\n\n#chatgpt #midjourney https://t.co/n5hlelzpty]
...	...	...
49991	i joined the @aipadtech x @moonsalecom exclusive giveaway for a chance to win being a part of the #aipad launch! get ready for a launch like no other!\n\ntg: https://t.co/5if0uosveq\n\n btceth #ai #artificialintelligence #crypto #chatgpt #future https://t.co/kp71xazn14	[i joined the @aipadtech x @moonsalecom exclusive giveaway for a chance to win being a part of the #aipad launch!, get ready for a launch like no other!, tg: https://t.co/5if0uosveq\n\n btceth #ai #artificialintelligence #crypto #chatgpt #future https://t.co/kp71xazn14]
49992	@iamjohnoliver i think chatgpt is taking the piss. #lastweektonight https://t.co/tvlmpntlbw	[@iamjohnoliver i think chatgpt is taking the piss., #lastweektonight https://t.co/tvlmpntlbw]
49993	digital marketers adopt new skills. artificial intelligence is completely going to capture your jobs. in a few months there will be no work for you in industry #chatgpt \n\nseo\nsmm\ncontent creators\ngraphic designer\nadvertising\nbpo/kpo\n\nabove 90% to 95% chances to drop out.	[digital marketers adopt new skills., artificial intelligence is completely going to capture your jobs., in a few months there will be no work for you in industry #chatgpt \n\nseo\nsmm\ncontent creators\ngraphic designer\nadvertising\nbpo/kpo\n\nabove 90% to 95% chances to drop out.]

Examples:

```
only_english_tweets.loc[49993:49995, ['Text', 'SentenceSegmentedTweet']]
```

	Text	SentenceSegmentedTweet
49993	digital marketers adopt new skills. artificial intelligence is completely going to capture your jobs. in a few months there will be no work for you in industry #chatgpt \n\nseo\nsmm\ncontent creators\ngraphic designer\nadvertising\nbpo/kpo\n\nabove 90% to 95% chances to drop out.	[digital marketers adopt new skills., artificial intelligence is completely going to capture your jobs., in a few months there will be no work for you in industry #chatgpt \n\nseo\nsmm\ncontent creators\ngraphic designer\nadvertising\nbpo/kpo\n\nabove 90% to 95% chances to drop out.]
49995	remember when @twitter was down a lot in the early years cause too many people were using it. @openai is having the same problem with #chatgpt now. a shift is here.	[remember when @twitter was down a lot in the early years cause too many people were using it., @openai is having the same problem with #chatgpt now., a shift is here.]

Removal of HTML Tags

```
only_english_tweets[['Source', 'SourceWithoutHTML']]
```

	Source	SourceWithoutHTML
1	Twitter for iPhone	Twitter for iPhone
3	Twitter for Android	Twitter for Android
6	Twitter Web App	Twitter Web App
7	Twitter Web App	Twitter Web App
8	Twitter for iPhone	Twitter for iPhone
...	...	...
49991	ViralSweep App	ViralSweep App
49992	Twitter Web App	Twitter Web App
49993	Twitter for Android	Twitter for Android
49995	Twitter for iPhone	Twitter for iPhone
49999	drumup.io	drumup.io

32076 rows × 2 columns

References:

1) https://rollbar.com/blog/python-recursionerror/#:~:text=Increasing%20the%20recursion%20limit%3A%20Python,recursion%20is%20not%20properly%20controlled

2) geeksforgeeks.com

3) https://www.nltk.org/_modules/nltk/tokenize/regexp.html

4) https://towardsdatascience.com/stemming-corpus-with-nltk-7a6a6d02d3e5

5) Canvas Class Slides: https://canvas.illinois.edu/courses/37566/pages/course-schedule

6) https://www.nltk.org/book/ch03.html

7) https://www.nltk.org/