- Detravious Jamari Brinkley
- HW1
- CSCI-544: Applied Natural Language Processing
- python version: 3.11.4

In [ ]:

```
import pandas as pd
import numpy as np
import nltk
nltk.download('wordnet')
import re
from bs4 import BeautifulSoup

from sklearn.feature_extraction.text import TfidfVectorizer

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

from nltk.stem import WordNetLemmatizer

from sklearn.model_selection import train_test_split

import sklearn
from sklearn.linear_model import Perceptron, LogisticRegression
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import MultinomialNB
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/brinkley97/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Read Data

In [ ]:

```
dataset = "../datasets/amazon_reviews_us_Office_Products_v1_00.tsv"
amazon_reviews_copy_df = pd.read_csv(dataset, sep='\t', on_bad_lines='skip', low_memory=False)
```

Keep Reviews and Ratings

In [ ]:
```
reviews_ratings_df = amazon_reviews_copy_df.loc[0:, ['star_rating', 'review_body']]
reviews_ratings_df.reset_index(drop=True)
```

Out[ ]:

	star_rating	review_body
0	5	Great product.
1	5	What's to say about this commodity item except...
2	5	Haven't used yet, but I am sure I will like it.
3	1	Although this was labeled as &#34;new&#34; the...
4	4	Gorgeous colors and easy to use
...	...	...
2640249	4	I can't live anymore whithout my Palm III. But...
2640250	4	Although the Palm Pilot is thin and compact it...
2640251	4	This book had a lot of great content without b...
2640252	5	I am teaching a course in Excel and am using t...
2640253	5	A very comprehensive layout of exactly how Vis...

2640254 rows × 2 columns

In [ ]:
```
reviews_ratings_df['review_body'].astype(str)
reviews_ratings_df
```

Out[ ]:

	star_rating	review_body
0	5	Great product.
1	5	What's to say about this commodity item except...
2	5	Haven't used yet, but I am sure I will like it.
3	1	Although this was labeled as &#34;new&#34; the...
4	4	Gorgeous colors and easy to use
...	...	...
2640249	4	I can't live anymore whithout my Palm III. But...
2640250	4	Although the Palm Pilot is thin and compact it...
2640251	4	This book had a lot of great content without b...
2640252	5	I am teaching a course in Excel and am using t...
2640253	5	A very comprehensive layout of exactly how Vis...

2640254 rows × 2 columns

In [ ]:
```
average_length_before_cleaning = reviews_ratings_df['review_body'][reviews_ratings_df['review_body'].apply(type)
print("Average length of the reviews in terms of character length BEFORE cleaning", average_length_before_cleani
```

```
Average length of the reviews in terms of character length BEFORE cleaning 285.2706194509257
```

In [ ]:
```
def generate_sample_reviews(df: pd.DataFrame, review_col_name: str, number_of_reviews: int = 3):
    """Include reviews and ratings

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    review_col_name: `str`
        The specific_column to get the reviews and ratings of

    number_of_reviews: `int`
        Number of samples to include
    Return
    ------
    Nothing; instead, print the reviews with ratings
    """


    columns_to_include = [review_col_name, 'star_rating']

    # Initialize an empty list to store dictionaries
    list_of_dicts = []

    # Iterate over the specified columns and retrieve the first three rows
    for row in df[columns_to_include].head(3).to_dict(orient='records'):
        list_of_dicts.append({'star_rating': row['star_rating'], review_col_name: row[review_col_name]})

    for dictionary in list_of_dicts:
        print(dictionary)
```

Select 100000 reviews randomly from positive and negative classes

In [ ]:

```
def update_data_type(df: pd.DataFrame, col_name: str):
    """Update the data type of the star ratings

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with rating values

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the new sentiment appened

    """

    valid_ratings = ['1','2','3','4','5']
    star_rating_series = df[col_name].copy()

    # Convert type to strings
    star_rating_series.astype('str')

    # Check valid list and see which of our stars match
    rows = star_rating_series.index
    is_rating_in_valid_ratings = rows[star_rating_series.isin(valid_ratings)]

    # Convert to list
    is_rating_in_valid_ratings = is_rating_in_valid_ratings.to_list()

    updated_df = df.iloc[is_rating_in_valid_ratings]
    return updated_df
```

In [ ]:

```
reviews_ratings_df = update_data_type(reviews_ratings_df, 'star_rating')
```

In [ ]:

```
reviews_ratings_df
```

Out[ ]:

	star_rating	review_body
0	5	Great product.
1	5	What's to say about this commodity item except...
2	5	Haven't used yet, but I am sure I will like it.
3	1	Although this was labeled as &#34;new&#34; the...
4	4	Gorgeous colors and easy to use
...	...	...
2640249	4	I can't live anymore whithout my Palm III. But...
2640250	4	Although the Palm Pilot is thin and compact it...
2640251	4	This book had a lot of great content without b...
2640252	5	I am teaching a course in Excel and am using t...
2640253	5	A very comprehensive layout of exactly how Vis...

2640237 rows × 2 columns

In [ ]:

```
print("# reviews per rating", reviews_ratings_df['star_rating'].value_counts())
```

```
# reviews per rating star_rating
5    1582812
4     418371
1     306979
3     193691
2     138384
Name: count, dtype: int64
```

In [ ]:

```python
def separate_reviews_by_rating(df: pd.DataFrame, rating_col: str, threshold: int, sentiment_type: str):
    """Categorizes reviews by adding a rating

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    rating_col: `str`
        Column with rating values

    threshold: `int`
        Where to split the ratings such that categories can be formed

    sentiment_type: `str`
        One of three types of sentiment: positive, negative, or neural

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the new sentiment appened
    """


    if sentiment_type == 'positive_sentiment':
        positive_review_threshold = df[rating_col].astype('int32') > threshold
        df = df[positive_review_threshold]
        df[sentiment_type] = 1

    elif sentiment_type == 'negative_sentiment':
        positive_review_threshold = df[rating_col].astype('int32') < threshold
        df = df[positive_review_threshold]
        df[sentiment_type] = 0

    elif sentiment_type == 'neutral_sentiment':
        positive_review_threshold = df[rating_col].astype('int32') == threshold
        df = df[positive_review_threshold]
        df[sentiment_type] = 3

    return df
```

In [ ]:

```
positive_sentiment_df = separate_reviews_by_rating(reviews_ratings_df, 'star_rating', 3, 'positive_sentiment')
positive_sentiment_df
```

```
/var/folders/fz/zn5r8vq12nv5p23dtlr15sk40000gn/T/ipykernel_11636/4050413545.py:28: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy
  df[sentiment_type] = 1
```

Out[ ]:

	star_rating	review_body	positive_sentiment
0	5	Great product.	1
1	5	What's to say about this commodity item except...	1
2	5	Haven't used yet, but I am sure I will like it.	1
4	4	Gorgeous colors and easy to use	1
5	5	Perfect for planning weekly meals. Removrd the...	1
...	...	...	...
2640249	4	I can't live anymore whithout my Palm III. But...	1
2640250	4	Although the Palm Pilot is thin and compact it...	1
2640251	4	This book had a lot of great content without b...	1
2640252	5	I am teaching a course in Excel and am using t...	1
2640253	5	A very comprehensive layout of exactly how Vis...	1

2001183 rows × 3 columns

In [ ]:

```
print("# positive sentiment: ", len(positive_sentiment_df))
print()
```

```
# positive sentiment:  2001183
```

In [ ]:

```
negative_sentiment_df = separate_reviews_by_rating(reviews_ratings_df, 'star_rating', 3, 'negative_sentiment')
negative_sentiment_df
```

```
/var/folders/fz/zn5r8vq12nv5p23dtlr15sk40000gn/T/ipykernel_11636/4050413545.py:33: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy
  df[sentiment_type] = 0
```

Out[ ]:

	star_rating	review_body	negative_sentiment
3	1	Although this was labeled as &#34;new&#34; the...	0
13	1	worked about a month then died	0
20	1	The phone did not work. No Dial Tone. Not wo...	0
27	1	Not laminated and no reinforced holes for hang...	0
28	1	Cartridge was over filled, black smears on pap...	0
...	...	...	...
2640139	2	This purchase was intended for a home office s...	0
2640149	2	I bought a Palm V from Amazon and thought it w...	0
2640151	1	The display is excellent - it's a good size an...	0
2640201	1	All the CE based hand held or palm computers h...	0
2640235	1	The Litium-ion batery failed from the start th...	0

445363 rows × 3 columns

In [ ]:

```
print("# negative sentiment: ", len(negative_sentiment_df))
print()
```

```
# negative sentiment:  445363
```

In [ ]:
```
neutral_sentiment_df = separate_reviews_by_rating(reviews_ratings_df, 'star_rating', 3, 'neutral_sentiment')
neutral_sentiment_df
```

```
/var/folders/fz/zn5r8vq12nv5p23dtlr15sk40000gn/T/ipykernel_11636/4050413545.py:38: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#retu
rning-a-view-versus-a-copy
  df[sentiment_type] = 3
```

Out[ ]:

	star_rating	review_body	neutral_sentiment
48	3	Nice quality. Happy with the item	3
64	3	The batch I had exploded all over when I tried...	3
95	3	It is ok, but considering the price plus shipp...	3
133	3	Delighted to receive a sample of these to try ...	3
145	3	I use this light in a dark area of my closet. ...	3
...	...	...	...
2640209	3	I was VERY disappointed to receive my Palm V a...	3
2640219	3	Very basic. The book spends a lot of time des...	3
2640225	3	Being a Newton devotee, switching to the Palm ...	3
2640234	3	I have a US Robotics Palm Pro (we go back a wa...	3
2640242	3	Bought Palm V and was disappointed to learn th...	3

193691 rows × 3 columns

In [ ]:
```
print("# neutral sentiment: ", len(neutral_sentiment_df))
print()
```

```
# neutral sentiment:  193691
```

In [ ]:
```
pos_rand_sampled_df = positive_sentiment_df.sample(100000)
pos_rand_sampled_df
```

Out[ ]:

	star_rating	review_body	positive_sentiment
2096055	5	I bought this item as a replacement to a TI-85...	1
380171	5	This is a great little printer from Epson, but...	1
632293	4	I've been looking for a front pocket wallet, a...	1
717003	4	It beats licking the envelopes. The sponge ti...	1
709281	5	I love this product. I have read many review ...	1
...	...	...	...
909182	5	Works well. Perfect for home use such as credi...	1
783087	5	Exactly what we wanted for our motor home. Wor...	1
1614362	5	Best ever!! Love the fact that you can print t...	1
64144	5	Perfect and fast shipping!	1
605991	5	Love love love my filofax. I use it as a purse...	1

100000 rows × 3 columns

In [ ]:

```python
neg_rand_sampled_df = negative_sentiment_df.sample(100000)
neg_rand_sampled_df
```

Out[ ]:

	star_rating	review_body	negative_sentiment
2087477	1	Yes they feel weird... like jelly (silicone) &...	0
2636736	1	The caller ID and answering machine worked ver...	0
2297691	1	I bought Royal rub ons at a Michaels store...t...	0
1521012	2	this stuff is just boring. You can not really...	0
1966489	1	Ink cartridges were recognized by printer as r...	0
...	...	...	...
86817	1	this is my second headset within 6 months. t...	0
1447712	1	These are not erasable, so they have ruined ou...	0
2200443	1	I purchased these phones two years ago. I was...	0
1122578	1	This does not work, planning to return for ref...	0
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0

100000 rows × 3 columns

In [ ]:

```
reviews_ratings_df = pd.concat([pos_rand_sampled_df, neg_rand_sampled_df])
reviews_ratings_df
```

Out[ ]:

	star_rating	review_body	positive_sentiment	negative_sentiment
2096055	5	I bought this item as a replacement to a TI-85...	1.0	NaN
380171	5	This is a great little printer from Epson, but...	1.0	NaN
632293	4	I've been looking for a front pocket wallet, a...	1.0	NaN
717003	4	It beats licking the envelopes. The sponge ti...	1.0	NaN
709281	5	I love this product. I have read many review ...	1.0	NaN
...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	NaN	0.0
1447712	1	These are not erasable, so they have ruined ou...	NaN	0.0
2200443	1	I purchased these phones two years ago. I was...	NaN	0.0
1122578	1	This does not work, planning to return for ref...	NaN	0.0
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	NaN	0.0

200000 rows × 4 columns

In [ ]:

```
pos_sentiment = reviews_ratings_df['positive_sentiment'].dropna()
pos_sentiment
```

Out[ ]:

```
2096055    1.0
380171     1.0
632293     1.0
717003     1.0
709281     1.0
          ...
909182     1.0
783087     1.0
1614362    1.0
64144      1.0
605991     1.0
Name: positive_sentiment, Length: 100000, dtype: float64
```

In [ ]:

```
neg_sentiment = reviews_ratings_df['negative_sentiment'].dropna()
neg_sentiment
```

```
Out[ ]: 2087477    0.0
        2636736    0.0
        2297691    0.0
        1521012    0.0
        1966489    0.0
                  ...
        86817      0.0
        1447712    0.0
        2200443    0.0
        1122578    0.0
        657766     0.0
        Name: negative_sentiment, Length: 100000, dtype: float64
```

```
In [ ]: reviews_ratings_df['sentiment'] = pd.concat([pos_sentiment, neg_sentiment])
```

```
In [ ]: reviews_ratings_df
```

Out[ ]:

	star_rating	review_body	positive_sentiment	negative_sentiment	sentiment
2096055	5	I bought this item as a replacement to a TI-85...	1.0	NaN	1.0
380171	5	This is a great little printer from Epson, but...	1.0	NaN	1.0
632293	4	I've been looking for a front pocket wallet, a...	1.0	NaN	1.0
717003	4	It beats licking the envelopes. The sponge ti...	1.0	NaN	1.0
709281	5	I love this product. I have read many review ...	1.0	NaN	1.0
...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	NaN	0.0	0.0
1447712	1	These are not erasable, so they have ruined ou...	NaN	0.0	0.0
2200443	1	I purchased these phones two years ago. I was...	NaN	0.0	0.0
1122578	1	This does not work, planning to return for ref...	NaN	0.0	0.0
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	NaN	0.0	0.0

200000 rows × 5 columns

```
In [ ]: reviews_sentiment_df = reviews_ratings_df.drop(columns=['positive_sentiment', 'negative_sentiment'])
        reviews_sentiment_df
```

Out[ ]:

	star_rating	review_body	sentiment
2096055	5	I bought this item as a replacement to a TI-85...	1.0
380171	5	This is a great little printer from Epson, but...	1.0
632293	4	I've been looking for a front pocket wallet, a...	1.0
717003	4	It beats licking the envelopes. The sponge ti...	1.0
709281	5	I love this product. I have read many review ...	1.0
...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0
1447712	1	These are not erasable, so they have ruined ou...	0.0
2200443	1	I purchased these phones two years ago. I was...	0.0
1122578	1	This does not work, planning to return for ref...	0.0
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0

200000 rows × 3 columns

In [ ]:

```
reviews_sentiment_df['review_body'].fillna(' ', inplace=True)
```

In [ ]:

```
print("Base review body:")
generate_sample_reviews(reviews_sentiment_df, 'review_body', 3)
```

```
Base review body:
{'star_rating': '5', 'review_body': "I bought this item as a replacement to a TI-85. I used that one because i l
iked the large screen and could see what i was typing (in the event you make a mistake). Although this TI-30X on
ly has a 2 line display, it's perfect and much cheaper. It also has the parenthesis buttons which makes combinin
g steps/formulas into a single entry...great for everyday..."}
{'star_rating': '5', 'review_body': "This is a great little printer from Epson, but has recently been replaced b
y the XP-420 model, the key difference being that the 320 has a 1.44 inch LCD screen, and the 420 has a larger
2.5 inch LCD screen.<br /><br />Otherwise all their stats are the same: 9/4.5 pages per minute (BW/color), 2400
pi, and print resolution up to 5760x1440. It scans to the usual formats: jpeg, tiff, PDF, png, etc. (Others too,
but those are the ones I use most.)<br /><br />As a Mac users I've always just downloaded the Epson drivers when
prompted, and in the case of the 320 setup was a breeze. Once setup, I could print from my Macbook Air over the
network, and from my iPad directly to the printer. Both devices found the printer easily and there were no glitc
hes. You can use a USB cable (not included), or print from an SD card, which is where the LCD comes in handy.<br
/><br />This is a compact printer, paper feeds in from the top (pull up the little tray), and while it says it c
an hold 100 sheets I didn't load that many--I suspect it would get pretty tight.<br /><br />The printer comes wi
th starter ink cartridges. Just enough to get you going. Of course, the printer is cheap and the ink will cost y
ou, but of course that's been the economy of printers for the past ten years or so: low cost machine, but they m
ake their money on the ink. That said, the quality of this printer is good, and for the price and quality of pri
nts (both paper and photo) it's a good value."}
{'star_rating': '4', 'review_body': "I've been looking for a front pocket wallet, and I decided to give this ite
m a shot.  There are some things I really like about it.  The industrial look and feel of it with the aluminum p
lates and the o-rings is very unique and attractive.  It holds all of my items securely and expands as needed.
The downside of it is that the aluminum plates are thicker than I expected, so it bulks up quicker than I had ho
ped."}
```

Data Cleaning

Lower case

In [ ]:

```python
def convert_reviews_to_lower_case(df: pd.DataFrame, col_name: str):
    """Convert all reviews to lower case

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    -------
    df: `pd.DataFrame`
        An updated DataFrame with the lower cased reviews
    """

    lower_case_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]
        # print(text_reviews_idx, type(text_review), text_review)

        # NOT all reviews are strings, thus all can't be converted to lower cased
        if type(text_review) != str:
            converted_str = str(text_review)
            # update_text_review = converted_str.lower()
            lower_case_reviews.append(text_review)
            # print(text_reviews_idx, update_text_review)
            # print()
        else:
            update_text_review = text_review.lower()
            lower_case_reviews.append(update_text_review)
            # print(text_reviews_idx, update_text_review)
            # print()

    updated_df['lower_cased'] = lower_case_reviews
    return updated_df
```

In [ ]:

```
reviews_lower_cased = convert_reviews_to_lower_case(reviews_sentiment_df, 'review_body')
```

In [ ]:

```
reviews_lower_cased
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...
...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...

200000 rows × 4 columns

In [ ]:

```
print("reviews_lower_cased:")
generate_sample_reviews(reviews_lower_cased, 'lower_cased', 3)
```

```
reviews_lower_cased:
{'star_rating': '5', 'lower_cased': "i bought this item as a replacement to a ti-85. i used that one because i l
iked the large screen and could see what i was typing (in the event you make a mistake). although this ti-30x on
ly has a 2 line display, it's perfect and much cheaper. it also has the parenthesis buttons which makes combinin
g steps/formulas into a single entry...great for everyday..."}
{'star_rating': '5', 'lower_cased': "this is a great little printer from epson, but has recently been replaced b
y the xp-420 model, the key difference being that the 320 has a 1.44 inch lcd screen, and the 420 has a larger
2.5 inch lcd screen.<br /><br />otherwise all their stats are the same: 9/4.5 pages per minute (bw/color), 2400
pi, and print resolution up to 5760x1440. it scans to the usual formats: jpeg, tiff, pdf, png, etc. (others too,
but those are the ones i use most.)<br /><br />as a mac users i've always just downloaded the epson drivers when
prompted, and in the case of the 320 setup was a breeze. once setup, i could print from my macbook air over the
network, and from my ipad directly to the printer. both devices found the printer easily and there were no glitc
hes. you can use a usb cable (not included), or print from an sd card, which is where the lcd comes in handy.<br
/><br />this is a compact printer, paper feeds in from the top (pull up the little tray), and while it says it c
an hold 100 sheets i didn't load that many--i suspect it would get pretty tight.<br /><br />the printer comes wi
th starter ink cartridges. just enough to get you going. of course, the printer is cheap and the ink will cost y
ou, but of course that's been the economy of printers for the past ten years or so: low cost machine, but they m
ake their money on the ink. that said, the quality of this printer is good, and for the price and quality of pri
nts (both paper and photo) it's a good value."}
{'star_rating': '4', 'lower_cased': "i've been looking for a front pocket wallet, and i decided to give this ite
m a shot.  there are some things i really like about it.  the industrial look and feel of it with the aluminum p
lates and the o-rings is very unique and attractive.  it holds all of my items securely and expands as needed.
the downside of it is that the aluminum plates are thicker than i expected, so it bulks up quicker than i had ho
ped."}
```

Remove HTML and URLs

In [ ]:

```python
def remove_html_and_urls(df: pd.DataFrame, col_name: str):
    """Remove HTML and URLs from all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the html_and_urls removed
    """

    # url_pattern = re.compile(r'https?://\S+|www\. \S+')

    cleaned_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]

        if isinstance(text_review, str):
            # Check and remove HTML tags
            has_html = bool(re.search('<.*?>', text_review))
            if has_html == True:
                # print("Review", text_reviews_idx, "has HTML -- ", text_review)
                pass

            no_html_review = re.sub('<.*?>', ' ', text_review)
            # print("Review", text_reviews_idx, "without HTML -- ", no_html_review)

            # Check and remove URLs
            has_url = bool(re.search(r'http\S+', no_html_review))
            if has_url == True:
                # print("Review", text_reviews_idx, "has URL --", no_html_review)
                pass

            no_html_url_review = re.sub(r'http\S+', '', no_html_review)
            # print("Review", text_reviews_idx, "without HTML, URL -- ", no_html_url_review)
            # print()
            cleaned_reviews.append(no_html_url_review)
        else:
            # print(text_reviews_idx, text_review)
            cleaned_reviews.append(text_review)


    updated_df['without_html_urls'] = cleaned_reviews
    return updated_df
```

In [ ]:

```
no_html_urls_df = remove_html_and_urls(reviews_lower_cased, 'lower_cased')
```

In [ ]:

```
no_html_urls_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...
...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...

200000 rows × 5 columns

In [ ]:

```
print("without_html_urls:")
generate_sample_reviews(no_html_urls_df, 'without_html_urls', 3)
```

```
without_html_urls:
{'star_rating': '5', 'without_html_urls': "i bought this item as a replacement to a ti-85. i used that one becau
se i liked the large screen and could see what i was typing (in the event you make a mistake). although this ti-
30x only has a 2 line display, it's perfect and much cheaper. it also has the parenthesis buttons which makes co
mbining steps/formulas into a single entry...great for everyday..."}
{'star_rating': '5', 'without_html_urls': "this is a great little printer from epson, but has recently been repl
aced by the xp-420 model, the key difference being that the 320 has a 1.44 inch lcd screen, and the 420 has a la
rger 2.5 inch lcd screen.  otherwise all their stats are the same: 9/4.5 pages per minute (bw/color), 2400 pi, a
nd print resolution up to 5760x1440. it scans to the usual formats: jpeg, tiff, pdf, png, etc. (others too, but
those are the ones i use most.)  as a mac users i've always just downloaded the epson drivers when prompted, and
in the case of the 320 setup was a breeze. once setup, i could print from my macbook air over the network, and f
rom my ipad directly to the printer. both devices found the printer easily and there were no glitches. you can u
se a usb cable (not included), or print from an sd card, which is where the lcd comes in handy.  this is a compa
ct printer, paper feeds in from the top (pull up the little tray), and while it says it can hold 100 sheets i di
dn't load that many--i suspect it would get pretty tight.  the printer comes with starter ink cartridges. just e
nough to get you going. of course, the printer is cheap and the ink will cost you, but of course that's been the
economy of printers for the past ten years or so: low cost machine, but they make their money on the ink. that s
aid, the quality of this printer is good, and for the price and quality of prints (both paper and photo) it's a
good value."}
{'star_rating': '4', 'without_html_urls': "i've been looking for a front pocket wallet, and i decided to give th
is item a shot.  there are some things i really like about it.  the industrial look and feel of it with the alum
inum plates and the o-rings is very unique and attractive.  it holds all of my items securely and expands as nee
ded.  the downside of it is that the aluminum plates are thicker than i expected, so it bulks up quicker than i
had hoped."}
```

Remove Contractions

In [ ]:

```
store_contractions = {
    "ain't": "am not",
    "aren't": "are not",
    "can't": "cannot",
    "couldn't": "could not",
    "didn't": "did not",
    "doesn't": "does not",
    "don't": "do not",
    "hadn't": "had not",
    "hasn't": "has not",
    "haven't": "have not",
    "he's": "he is",
    "isn't": "is not",
    "it's": "it is",
    "let's": "let us",
    "mustn't": "must not",
    "shan't": "shall not",
    "she's": "she is",
    "shouldn't": "should not",
    "that's": "that is",
    "there's": "there is",
    "they're": "they are",
    "wasn't": "was not",
    "we're": "we are",
    "weren't": "were not",
    "won't": "will not",
    "wouldn't": "would not",
    "you're": "you are",
    "you'll": "you will",
    "you'd": "you would",
    "we'll": "we will",
    "we've": "we have",
    "we'd": "we would",
    "I'm": "I am",
    "i've": "I have",
    "I've": "I have",
    "I'd": "I would",
    "it'll": "it will",
    "they'll": "they will",
    "they've": "they have",
    "they'd": "they would",
    "he'll": "he will",
    "he'd": "he would",
    "she'll": "she will",
    "we'd": "we would",
    "we'll": "we will",
    "you've": "you have",
    "you'd": "you would",
    "you'll": "you will",
    "I'll": "I will",
    "I'd": "I would",
    "it's": "it is",
    "it'd": "it would",
    "i'm": "I am",
    "he's": "he is",
    "he'll": "he will",
    "she's": "she is",
    "she'll": "she will",
    "we're": "we are",
    "we've": "we have",
    "we'll": "we will",
    "you're": "you are",
    "you've": "you have",
    "you'll": "you will",
    "they're": "they are",
    "they've": "they have",
    "they'll": "they will",
    "that's": "that is",
    "that'll": "that will",
    "that'd": "that would",
    "who's": "who is",
    "who'll": "who will",
    "who'd": "who would",
    "what's": "what is",
    "what'll": "what will",
    "what'd": "what would",
    "when's": "when is",
    "when'll": "when will",
    "when'd": "when would",
    "where's": "where is",
    "where'll": "where will",
    "where'd": "where would",
    "why's": "why is",
    "why'll": "why will",
    "why'd": "why would",
    "how's": "how is",
    "how'll": "how will",
    "how'd": "how would"
}
```

In [ ]:

```
def locate_and_replace_contractions(review):
    """Find the contractions to replace from a specific review

    Parameters
    ----------
    review: `str`
        A specific review

    Return
    ------
    non_contraction_review: `str`
        The updated specific review with contractions expanded
    """
    if isinstance(review, str):
        get_words = review.split()

        store_non_contraction_words = []

        for word in get_words:
            if word in store_contractions:
                non_contraction_form = store_contractions[word]
                # print(word, "-->", non_contraction_form)

                store_non_contraction_words.append(non_contraction_form)

            else:
                # print(word)
                store_non_contraction_words.append(word)

        non_contraction_review = ' '.join(store_non_contraction_words)
        return non_contraction_review
    else:
        return review
```

In [ ]:

```
def remove_contractions(df:pd.DataFrame, col_name: str):
    """Remove contractions from all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the extra spaces removed
    """

    without_contractions_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values
    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]

        # print("Review", text_reviews_idx, "with possible contraction(s) -- ", text_review)

        without_contraction = locate_and_replace_contractions(text_review)

        # print("Review", text_reviews_idx, "without contraction -- ", without_contraction)
        # print()

        without_contractions_reviews.append(without_contraction)

    updated_df['without_contractions'] = without_contractions_reviews
    return updated_df
```

In [ ]:

```
no_contractions_df = remove_contractions(no_html_urls_df, 'without_html_urls')
```

In [ ]:

```
no_contractions_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson, but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...	I have been looking for a front pocket wallet,...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge tip...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...	i love this product. i have read many review o...
...	...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...	this is my second headset within 6 months. the...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was ...
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work, planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...

200000 rows × 6 columns

In [ ]:

```
print("without_contractions:")
generate_sample_reviews(no_contractions_df, 'without_contractions', 3)
```

```
without_contractions:
{'star_rating': '5', 'without_contractions': 'i bought this item as a replacement to a ti-85. i used that one be
cause i liked the large screen and could see what i was typing (in the event you make a mistake). although this
ti-30x only has a 2 line display, it is perfect and much cheaper. it also has the parenthesis buttons which make
s combining steps/formulas into a single entry...great for everyday...'}
{'star_rating': '5', 'without_contractions': 'this is a great little printer from epson, but has recently been r
eplaced by the xp-420 model, the key difference being that the 320 has a 1.44 inch lcd screen, and the 420 has a
larger 2.5 inch lcd screen. otherwise all their stats are the same: 9/4.5 pages per minute (bw/color), 2400 pi,
and print resolution up to 5760x1440. it scans to the usual formats: jpeg, tiff, pdf, png, etc. (others too, but
those are the ones i use most.) as a mac users I have always just downloaded the epson drivers when prompted, an
d in the case of the 320 setup was a breeze. once setup, i could print from my macbook air over the network, and
from my ipad directly to the printer. both devices found the printer easily and there were no glitches. you can
use a usb cable (not included), or print from an sd card, which is where the lcd comes in handy. this is a compa
ct printer, paper feeds in from the top (pull up the little tray), and while it says it can hold 100 sheets i di
d not load that many--i suspect it would get pretty tight. the printer comes with starter ink cartridges. just e
nough to get you going. of course, the printer is cheap and the ink will cost you, but of course that is been th
e economy of printers for the past ten years or so: low cost machine, but they make their money on the ink. that
said, the quality of this printer is good, and for the price and quality of prints (both paper and photo) it is
a good value.'}
{'star_rating': '4', 'without_contractions': 'I have been looking for a front pocket wallet, and i decided to gi
ve this item a shot. there are some things i really like about it. the industrial look and feel of it with the a
luminum plates and the o-rings is very unique and attractive. it holds all of my items securely and expands as n
eeded. the downside of it is that the aluminum plates are thicker than i expected, so it bulks up quicker than i
had hoped.'}
```

Remove Non-alphabetical characters

In [ ]:

```
def remove_non_alphabetical_characters(df:pd.DataFrame, col_name: str):
    """Remove Non-alphabetical characters from all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the non-alphabetical characters removed
    """

    alphabetical_char_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values
    # print(text_reviews)

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]

        if isinstance(text_review, str):

            # Check for non-alphabetical characters
            has_non_alphabetical_char = bool(re.search(r'[^a-zA-Z]', text_review))
            if has_non_alphabetical_char == True:
                # print("Review", text_reviews_idx, "has HTML -- ", text_review)
                pass

            # Remove non-alphabetical characters
            with_alphabetical_char = re.sub(r'[^a-zA-Z\s]', ' ', text_review)
            # print("Review", text_reviews_idx, "has HTML -- ", with_alphabetical_char)
            alphabetical_char_reviews.append(with_alphabetical_char)
        else:
            alphabetical_char_reviews.append(text_review)

    updated_df['with_alpha_chars_only'] = alphabetical_char_reviews
    return updated_df
```

In [ ]:

```
only_alpha_chars_df = remove_non_alphabetical_characters(no_contractions_df, 'without_contractions')
```

In [ ]:

```
only_alpha_chars_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti ...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...	I have been looking for a front pocket wallet,...	I have been looking for a front pocket wallet ...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge tip...	it beats licking the envelopes the sponge tip...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...	i love this product. i have read many review o...	i love this product i have read many review o...
...	...	...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...	this is my second headset within 6 months. the...	this is my second headset within months the...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was ...	i purchased these phones two years ago i was ...

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm a few years ...

200000 rows × 7 columns

```
In [ ]: print("with_alpha_chars_only:")
        generate_sample_reviews(only_alpha_chars_df, 'with_alpha_chars_only', 3)
```

```
with_alpha_chars_only:
{'star_rating': '5', 'with_alpha_chars_only': 'i bought this item as a replacement to a ti     i used that one b
ecause i liked the large screen and could see what i was typing  in the event you make a mistake   although this
ti   x only has a   line display  it is perfect and much cheaper  it also has the parenthesis buttons which make
s combining steps formulas into a single entry   great for everyday   '}
{'star_rating': '5', 'with_alpha_chars_only': 'this is a great little printer from epson  but has recently been
replaced by the xp     model  the key difference being that the     has a      inch lcd screen  and the     has
a larger     inch lcd screen  otherwise all their stats are the same        pages per minute  bw color        pi
and print resolution up to     x      it scans to the usual formats  jpeg  tiff  pdf  png  etc   others too  but
those are the ones i use most   as a mac users I have always just downloaded the epson drivers when prompted  an
d in the case of the     setup was a breeze  once setup  i could print from my macbook air over the network  and
from my ipad directly to the printer  both devices found the printer easily and there were no glitches  you can
use a usb cable  not included   or print from an sd card  which is where the lcd comes in handy  this is a compa
ct printer  paper feeds in from the top  pull up the little tray   and while it says it can hold     sheets i di
d not load that many  i suspect it would get pretty tight  the printer comes with starter ink cartridges  just e
nough to get you going  of course  the printer is cheap and the ink will cost you  but of course that is been th
e economy of printers for the past ten years or so  low cost machine  but they make their money on the ink  that
said  the quality of this printer is good  and for the price and quality of prints  both paper and photo  it is
a good value '}
{'star_rating': '4', 'with_alpha_chars_only': 'I have been looking for a front pocket wallet  and i decided to g
ive this item a shot  there are some things i really like about it  the industrial look and feel of it with the
aluminum plates and the o rings is very unique and attractive  it holds all of my items securely and expands as
needed  the downside of it is that the aluminum plates are thicker than i expected  so it bulks up quicker than
i had hoped '}
```

Remove extra spaces

In [ ]:

```
def remove_extra_spaces(df:pd.DataFrame, col_name: str):
    """Remove extra spaces from all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the extra spaces removed
    """

    single_spaced_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values
    # print(text_reviews)

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]

        if isinstance(text_review, str):
        # Check if there are any extra spaces
            has_extra_space = bool(re.search(r' +', text_review))
            if has_extra_space == True:
                # print("Review", text_reviews_idx, "has extra space -- ", text_review)
                pass

            # Remove extra spaces
            single_spaced_review = re.sub(r' +', ' ', text_review)
            # print("Review", text_reviews_idx, "without extra space -- ", single_spaced_review)
            # print()

            single_spaced_reviews.append(single_spaced_review)
        else:
            single_spaced_reviews.append(text_review)
    updated_df['without_extra_space'] = single_spaced_reviews
    return updated_df
```

In [ ]:

```
no_extra_space_df = remove_extra_spaces(only_alpha_chars_df, 'with_alpha_chars_only')
```

In [ ]:

```
no_extra_space_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti ...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...	I have been looking for a front pocket wallet,...	I have been looking for a front pocket wallet ...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge tip...	it beats licking the envelopes the sponge tip...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...	i love this product. i have read many review o...	i love this product i have read many review o...
...	...	...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...	this is my second headset within 6 months. the...	this is my second headset within months the...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was ...	i purchased these phones two years ago i was ...

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm a few years ...

200000 rows × 8 columns

```
In [ ]: print("without_extra_space:")
        generate_sample_reviews(no_extra_space_df, 'without_extra_space', 3)
```

```
without_extra_space:
{'star_rating': '5', 'without_extra_space': 'i bought this item as a replacement to a ti i used that one because
i liked the large screen and could see what i was typing in the event you make a mistake although this ti x only
has a line display it is perfect and much cheaper it also has the parenthesis buttons which makes combining step
s formulas into a single entry great for everyday '}
{'star_rating': '5', 'without_extra_space': 'this is a great little printer from epson but has recently been rep
laced by the xp model the key difference being that the has a inch lcd screen and the has a larger inch lcd scre
en otherwise all their stats are the same pages per minute bw color pi and print resolution up to x it scans to
the usual formats jpeg tiff pdf png etc others too but those are the ones i use most as a mac users I have alway
s just downloaded the epson drivers when prompted and in the case of the setup was a breeze once setup i could p
rint from my macbook air over the network and from my ipad directly to the printer both devices found the printe
r easily and there were no glitches you can use a usb cable not included or print from an sd card which is where
the lcd comes in handy this is a compact printer paper feeds in from the top pull up the little tray and while i
t says it can hold sheets i did not load that many i suspect it would get pretty tight the printer comes with st
arter ink cartridges just enough to get you going of course the printer is cheap and the ink will cost you but o
f course that is been the economy of printers for the past ten years or so low cost machine but they make their
money on the ink that said the quality of this printer is good and for the price and quality of prints both pape
r and photo it is a good value '}
{'star_rating': '4', 'without_extra_space': 'I have been looking for a front pocket wallet and i decided to give
this item a shot there are some things i really like about it the industrial look and feel of it with the alumin
um plates and the o rings is very unique and attractive it holds all of my items securely and expands as needed
the downside of it is that the aluminum plates are thicker than i expected so it bulks up quicker than i had hop
ed '}
```

In [ ]:

```
average_length_after_cleaning = no_extra_space_df['review_body'][no_extra_space_df['review_body'].apply(type) ==
print("Average length of the reviews in terms of character length AFTER cleaning", average_length_after_cleaning
```

```
Average length of the reviews in terms of character length AFTER cleaning 317.42962
```

Pre-processing

remove the stop words

In [ ]:

```
def filter_stop_words(df:pd.DataFrame, col_name: str):
    """Filter stop words out from all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the extra spaces removed
    """

    without_stop_words_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values

    stop_words = set(stopwords.words("english"))

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]

        if isinstance(text_review, str):
            text_review_words = word_tokenize(text_review)
            # print("Before stop word removal", text_reviews_idx, " -- ", text_review)

            filtered_review = []

            for text_review_words_idx in range(len(text_review_words)):
                text_review_word = text_review_words[text_review_words_idx]

                # Check if review word is a stop word
                if text_review_word in stop_words:
                    # print("  Stop word -- ", text_review_word)
                    pass
                else:
                    # print(text_review_word, " -- is NOT a stop word in review")
                    filtered_review.append(text_review_word)


            filtered_review = " ".join(filtered_review)
            # print("After stop word removal", text_reviews_idx, " -- ", filtered_review)
            # print()

            without_stop_words_reviews.append(filtered_review)
        else:
            without_stop_words_reviews.append(text_review)


    updated_df['without_stop_words'] = without_stop_words_reviews
    return updated_df
```

In [ ]:

```
no_stop_words_df = filter_stop_words(no_extra_space_df, 'without_extra_space')
```

In [ ]:

```
no_stop_words_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti ...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...	I have been looking for a front pocket wallet,...	I have been looking for a front pocket wallet ...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge tip...	it beats licking the envelopes the sponge tip...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...	i love this product. i have read many review o...	i love this product i have read many review o...
...	...	...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...	this is my second headset within 6 months. the...	this is my second headset within months the...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was ...	i purchased these phones two years ago i was ...

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm a few years ...

200000 rows × 9 columns

```
In [ ]: print("without_stop_words:")
        generate_sample_reviews(no_stop_words_df, 'without_stop_words', 3)
```

```
without_stop_words:
{'star_rating': '5', 'without_stop_words': 'bought item replacement ti used one liked large screen could see typ
ing event make mistake although ti x line display perfect much cheaper also parenthesis buttons makes combining
steps formulas single entry great everyday'}
{'star_rating': '5', 'without_stop_words': 'great little printer epson recently replaced xp model key difference
inch lcd screen larger inch lcd screen otherwise stats pages per minute bw color pi print resolution x scans usu
al formats jpeg tiff pdf png etc others ones use mac users I always downloaded epson drivers prompted case setup
breeze setup could print macbook air network ipad directly printer devices found printer easily glitches use usb
cable included print sd card lcd comes handy compact printer paper feeds top pull little tray says hold sheets l
oad many suspect would get pretty tight printer comes starter ink cartridges enough get going course printer che
ap ink cost course economy printers past ten years low cost machine make money ink said quality printer good pri
ce quality prints paper photo good value'}
{'star_rating': '4', 'without_stop_words': 'I looking front pocket wallet decided give item shot things really l
ike industrial look feel aluminum plates rings unique attractive holds items securely expands needed downside al
uminum plates thicker expected bulks quicker hoped'}
```

perform lemmatization

- "A sentence with many words"
    - "words" -> word

In [ ]:

```
def lemmentize_review(df:pd.DataFrame, col_name: str):
    """Lemmentize all reviews

    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    df: `pd.DataFrame`
        An updated DataFrame with the extra spaces removed
    """

    lemmed_reviews = []
    updated_df = df.copy()
    text_reviews = df[col_name].values

    lem = WordNetLemmatizer()

    for text_reviews_idx in range(len(text_reviews)):
        text_review = text_reviews[text_reviews_idx]
        if isinstance(text_review, str):
            words_in_review = word_tokenize(text_review)

            # print("Before lem update", text_reviews_idx, " -- ", text_review)
            # print("Lemmed words", words_in_review)


            lemmed_sentence = []

            # Split review into words
            for lemmed_words_idx in range(len(words_in_review)):
                word = words_in_review[lemmed_words_idx]

                apply_lemmatization = lem.lemmatize(word)
                # print(apply_lemmatization)

                lemmed_sentence.append(apply_lemmatization)
                filtered_review = " ".join(lemmed_sentence)
            # print("After lem update -- ", filtered_review)
            # print()

            lemmed_reviews.append(filtered_review)
        else:
            lemmed_reviews.append(text_review)

    updated_df['lemmed_reviews'] = lemmed_reviews
    return updated_df
```

In [ ]:
```
lemmed_df = lemmentize_review(no_stop_words_df, 'without_stop_words')
```

In [ ]:
```
lemmed_df
```

Out[ ]:

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
2096055	5	I bought this item as a replacement to a TI-85...	1.0	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti-85...	i bought this item as a replacement to a ti ...
380171	5	This is a great little printer from Epson, but...	1.0	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson, but...	this is a great little printer from epson but...
632293	4	I've been looking for a front pocket wallet, a...	1.0	i've been looking for a front pocket wallet, a...	i've been looking for a front pocket wallet, a...	I have been looking for a front pocket wallet,...	I have been looking for a front pocket wallet ...
717003	4	It beats licking the envelopes. The sponge ti...	1.0	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge ti...	it beats licking the envelopes. the sponge tip...	it beats licking the envelopes the sponge tip...
709281	5	I love this product. I have read many review ...	1.0	i love this product. i have read many review ...	i love this product. i have read many review ...	i love this product. i have read many review o...	i love this product i have read many review o...
...	...	...	...	...	...	...	...
86817	1	this is my second headset within 6 months. t...	0.0	this is my second headset within 6 months. t...	this is my second headset within 6 months. t...	this is my second headset within 6 months. the...	this is my second headset within months the...
1447712	1	These are not erasable, so they have ruined ou...	0.0	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable, so they have ruined ou...	these are not erasable so they have ruined ou...
2200443	1	I purchased these phones two years ago. I was...	0.0	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was...	i purchased these phones two years ago. i was ...	i purchased these phones two years ago i was ...

	star_rating	review_body	sentiment	lower_cased	without_html_urls	without_contractions	with_alpha_chars_only
1122578	1	This does not work, planning to return for ref...	0.0	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work, planning to return for ref...	this does not work planning to return for ref...
657766	1	I bought a Plustek Opticfilm 7300 a few years ...	0.0	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm 7300 a few years ...	i bought a plustek opticfilm a few years ...

200000 rows × 10 columns

In [ ]:
```
print("without_stop_words:")
generate_sample_reviews(lemmed_df, 'lemmed_reviews', 3)
```

```
without_stop_words:
{'star_rating': '5', 'lemmed_reviews': 'bought item replacement ti used one liked large screen could see typing 
event make mistake although ti x line display perfect much cheaper also parenthesis button make combining step f
ormula single entry great everyday'}
{'star_rating': '5', 'lemmed_reviews': 'great little printer epson recently replaced xp model key difference inc
h lcd screen larger inch lcd screen otherwise stats page per minute bw color pi print resolution x scan usual fo
rmat jpeg tiff pdf png etc others one use mac user I always downloaded epson driver prompted case setup breeze s
etup could print macbook air network ipad directly printer device found printer easily glitch use usb cable incl
uded print sd card lcd come handy compact printer paper feed top pull little tray say hold sheet load many suspe
ct would get pretty tight printer come starter ink cartridge enough get going course printer cheap ink cost cour
se economy printer past ten year low cost machine make money ink said quality printer good price quality print p
aper photo good value'}
{'star_rating': '4', 'lemmed_reviews': 'I looking front pocket wallet decided give item shot thing really like i
ndustrial look feel aluminum plate ring unique attractive hold item securely expands needed downside aluminum pl
ate thicker expected bulk quicker hoped'}
```

TF-IDF Feature Extraction

In [ ]:
```
def tf_idf_feature_extraction(df: pd.DataFrame, col_name: str):
    """Extract the TF-IDF features from the reviews.
    Parameters
    ----------
    df: `pd.DataFrame`
        The data

    col_name: `str`
        Column with reviews

    Return
    ------
    tf_idf_features:
        A matrix containing the TF-IDF features extracted

    """

    vectorizer = TfidfVectorizer()
    tf_idf_features = vectorizer.fit_transform(df[col_name])

    return tf_idf_features
```

```
In [ ]: tf_idf_features = tf_idf_feature_extraction(lemmed_df, 'lemmed_reviews')
```

```
In [ ]: tf_idf_features[0]
```

```
Out[ ]: <1x56557 sparse matrix of type '<class 'numpy.float64'>'
        	with 31 stored elements in Compressed Sparse Row format>
```

Split Features and Sentiment Labels

```
In [ ]: sentiments = lemmed_df['sentiment']
        sentiments.shape
```

```
Out[ ]: (200000,)
```

```
In [ ]: X_train, X_test, y_train, y_test = train_test_split(tf_idf_features, sentiments, test_size=0.2, random_state=42)
        X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
Out[ ]: ((160000, 56557), (40000, 56557), (160000,), (40000,))
```

Models + Evaluation Metrics

In [ ]:

```
def eval_accuracy(y_true, y_prediction):
    return sklearn.metrics.accuracy_score(y_true, y_prediction)

def eval_precision(y_true, y_prediction):
    return sklearn.metrics.precision_score(y_true, y_prediction)

def eval_recall(y_true, y_prediction):
    return sklearn.metrics.recall_score(y_true, y_prediction)

def eval_f1_score(y_true, y_prediction):
    return sklearn.metrics.f1_score(y_true, y_prediction)
```

In [ ]:

```
def train_eval_metric(y_train_true, y_train_predictions):
    accuracy = eval_accuracy(y_train_true, y_train_predictions)
    precision = eval_precision(y_train_true, y_train_predictions)
    recall = eval_recall(y_train_true, y_train_predictions)
    f1 = eval_f1_score(y_train_true, y_train_predictions)

    metrics_dict = {
        'Accuracy': accuracy,
        'Precision': precision,
        'Recall': recall,
        'F1 Score': f1
    }

    return metrics_dict

def test_eval_metric(y_test_true, y_test_predictions):
    accuracy = eval_accuracy(y_test_true, y_test_predictions)
    precision = eval_precision(y_test_true, y_test_predictions)
    recall = eval_recall(y_test_true, y_test_predictions)
    f1 = eval_f1_score(y_test_true, y_test_predictions)

    metrics_dict = {
        'Accuracy': accuracy,
        'Precision': precision,
        'Recall': recall,
        'F1 Score': f1
    }

    return metrics_dict
```

Perceptron

In [ ]:

```python
def perceptron_model(X_train, X_test, y_train, y_test):

    technique = Perceptron(tol=1e-3, random_state=0)
    technique.fit(X_train, y_train)
    y_train_predictions = technique.predict(X_train)
    y_test_predictions = technique.predict(X_test)


    train_metrics = train_eval_metric(y_train, y_train_predictions)
    test_metrics = test_eval_metric(y_test, y_test_predictions)

    return train_metrics, test_metrics
```

In [ ]:

```python
perceptron_train_metrics, perceptron_test_metrics = perceptron_model(X_train, X_test, y_train, y_test)
```

In [ ]:

```python
perceptron_train_metrics, perceptron_test_metrics
```

Out[ ]:

```
({'Accuracy': 0.89669375,
  'Precision': 0.9097575460249425,
  'Recall': 0.8807729323684178,
  'F1 Score': 0.8950306417299082},
 {'Accuracy': 0.850075,
  'Precision': 0.8652019622168876,
  'Recall': 0.8292402340819287,
  'F1 Score': 0.8468394841016473})
```

SVM

In [ ]:

```python
def svm_model(X_train, X_test, y_train, y_test):

    technique = LinearSVC(tol=1e-3, random_state=0)
    technique.fit(X_train, y_train)
    y_train_predictions = technique.predict(X_train)
    y_test_predictions = technique.predict(X_test)
    train_metrics = train_eval_metric(y_train, y_train_predictions)
    test_metrics = test_eval_metric(y_test, y_test_predictions)

    return train_metrics, test_metrics
```

In [ ]:

```
svm_train_metrics, svm_test_metrics = svm_model(X_train, X_test, y_train, y_test)
```

```
/usr/local/lib/python3.11/site-packages/sklearn/svm/_classes.py:31: FutureWarning: The default value of `dual` w
ill change from `True` to `'auto'` in 1.5. Set the value of `dual` explicitly to suppress the warning.
  warnings.warn(
```

In [ ]:

```
svm_train_metrics, svm_test_metrics
```

Out[ ]:

```
({'Accuracy': 0.9305875,
  'Precision': 0.9318999561211058,
  'Recall': 0.929081205394528,
  'F1 Score': 0.9304884460356008},
 {'Accuracy': 0.8926,
  'Precision': 0.8939416754504844,
  'Recall': 0.8908117841244435,
  'F1 Score': 0.8923739853692755})
```

Logistic Regression

In [ ]:

```
def logistic_regression_model(X_train, X_test, y_train, y_test):

    technique = LogisticRegression(random_state=0)
    technique.fit(X_train, y_train)
    y_train_predictions = technique.predict(X_train)
    y_test_predictions = technique.predict(X_test)


    train_metrics = train_eval_metric(y_train, y_train_predictions)
    test_metrics = test_eval_metric(y_test, y_test_predictions)

    return train_metrics, test_metrics
```

In [ ]:

```
logistic_regression_train_metrics, logistic_regression_test_metrics = logistic_regression_model(X_train, X_test,
```

In [ ]:

```
logistic_regression_train_metrics, logistic_regression_test_metrics
```

Out[ ]:

```
({'Accuracy': 0.909425,
  'Precision': 0.9123604274978285,
  'Recall': 0.9058832352169185,
  'F1 Score': 0.9091102943943404},
 {'Accuracy': 0.8961,
  'Precision': 0.897295670061713,
  'Recall': 0.8945130795778522,
  'F1 Score': 0.8959022142069933})
```

Naive Bayes

In [ ]:

```
def naive_bayes_model(X_train, X_test, y_train, y_test):

    technique = MultinomialNB()
    technique.fit(X_train.toarray(), y_train)
    y_train_predictions = technique.predict(X_train)
    y_test_predictions = technique.predict(X_test)

    train_metrics = train_eval_metric(y_train, y_train_predictions)
    test_metrics = test_eval_metric(y_test, y_test_predictions)

    return train_metrics, test_metrics
```

In [ ]:

```
naive_bayes_train_metrics, naive_bayes_test_metrics = naive_bayes_model(X_train, X_test, y_train, y_test)
```

In [ ]:

```
naive_bayes_train_metrics, naive_bayes_test_metrics
```

Out[ ]:

```
({'Accuracy': 0.876,
  'Precision': 0.8845667097038107,
  'Recall': 0.8648868224030397,
  'F1 Score': 0.8746160749270069},
 {'Accuracy': 0.860275,
  'Precision': 0.8674864782120625,
  'Recall': 0.8503476216675837,
  'F1 Score': 0.8588315526255967})
```

In [ ]:

In [ ]: