IST664

NATURAL LANGUAGE PROCESSING

FALL 2023

HOMEWORK 1 - CORPUS STATISTICS AND PYTHON PROGRAMMING

HENDI KUSHTA

Problem statement

The spread of fake news has become a serious problem in today's digital age and is undermining confidence in online information sources. By utilizing Natural Language Processing (NLP) methods to examine a dataset made up of both fake and authentic news stories, this study seeks to address this problem. The objective is to use methods we have learned in class to gain understanding of these stories and improve our capacity to tell fact from fiction in news.

Preprocessing of data is one main goal. We want to extract textual information from both CSV files using Python and perform the necessary preprocessing operations individually. Key choices include tokenization, case handling, stop word usage, and possible lemmatization; our Jupyter Notebook includes justifications for each.

Then, we dive into data analysis to develop a thorough comprehension of textual content:

- Identifying the top 50 stop words by frequency in both fake and real news articles.
- Compiling the top 50 content words by frequency in both fake and real news articles.
- Extracting the top 50 bigrams by frequency in both fake and real news articles.
- Evaluating the top 50 bigrams by their Mutual Information scores (with a minimum frequency of 5) in both fake and real news articles.
- Identifying the top 50 adjective words in both fake and real news articles.

Additionally, we calculate various textual statistics for each article, such as the total number of words, content words, capitalized words (excluding "I"), exclamation marks, and punctuation marks. This information enriches the dataset for further analysis.

The interpretation of results is the main goal. We compare analysis results from real and fake news stories in order to find similarities and differences. Do these papers, for instance, have a comparable number of capitalized words? Which bigrams or content words fall under which group more frequently? These observations support broader initiatives to eliminate false information and promote a more critical online information landscape.

Data

The "ISOT Fake News Dataset," which includes over 12,600 stories, is the dataset under discussion. It includes both legitimate news stories taken from Reuters.com and false news pieces from dubious websites that have been exposed by trustworthy fact-checking groups.

There are two CSV files that make up the dataset: "True.csv" and "Fake.csv." While "Fake.csv" has more than 12,600 articles from various phony news sources, "True.csv" contains over 12,600 articles sourced from Reuters.com. The dataset contains the following details for each entry: the article's title, text, article type (genuine or fraudulent), and the publication date.

The focus of data collecting mostly covered the years 2016 to 2017 in order to coincide with information gathered for Kaggle.com. It's vital to note that the punctuation and typos found in the phony news items were not removed from the text even though the data was cleaned and processed.

Dataset Categories and Distribution:

Real-News

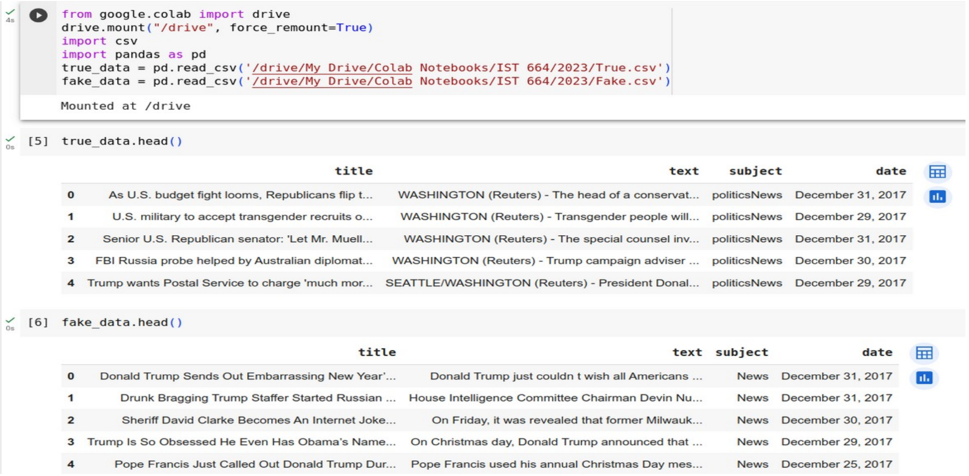
Number of Articles: 21,417
Categories:
World-News: 1,014 articles
Politics-News: 11,272 articles

Fake-News

Number of Articles: 23,481
Categories:
Government-News: 1,570 articles
Middle-East: 778 articles
US-News: 783 articles
Left-news: 4,459 articles
Politics: 6,841 articles
News: 9,050 articles

Data Reading/Loading

I am using pandas dataframe to read both csv files, as shown in the screen shot below.



```
from google.colab import drive
drive.mount("/drive", force_remount=True)
import csv
import pandas as pd
true_data = pd.read_csv('/drive/My Drive/Colab Notebooks/IST 664/2023/True.csv')
fake_data = pd.read_csv('/drive/My Drive/Colab Notebooks/IST 664/2023/Fake.csv')
```

```
Mounted at /drive
```

```
[5] true_data.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

```
[6] fake_data.head()
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

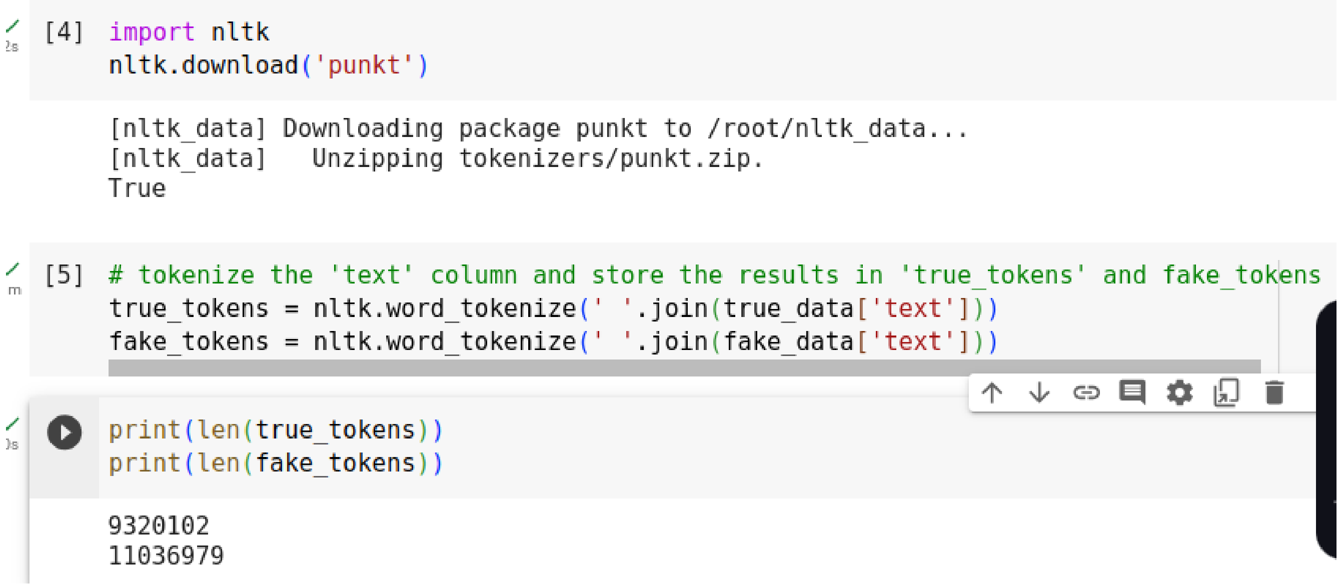
As we can see from the screen shot above, there are 4 columns for each dataset(title, text,subject, date).

Data Preprocessing

We must prepare our data before we can analyze both fake and true news stories. Similar to organizing and cleaning the components before cooking, this procedure is known as data preprocessing. In simple terms, it refers to improving the text's clarity and computer understanding. The text will be broken up into words, all capitalization will be in lowercase, numerals and punctuation will be removed, and only the most crucial words will be retained.

1. Tokenization

Tokenization, which is an essential NLP process, divides a text passage into smaller parts known as tokens. Tokens stand for specific words or phrases, which can make for more insightful analytical units than raw text.



```
[4] import nltk
    nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

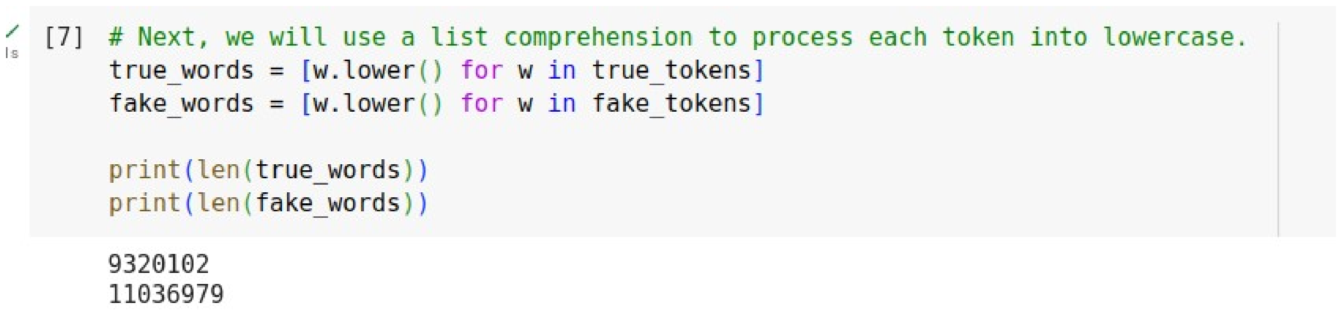
```
[5] # tokenize the 'text' column and store the results in 'true_tokens' and fake_tokens
    true_tokens = nltk.word_tokenize(' '.join(true_data['text']))
    fake_tokens = nltk.word_tokenize(' '.join(fake_data['text']))
```

```
print(len(true_tokens))
print(len(fake_tokens))
```

```
9320102
11036979
```

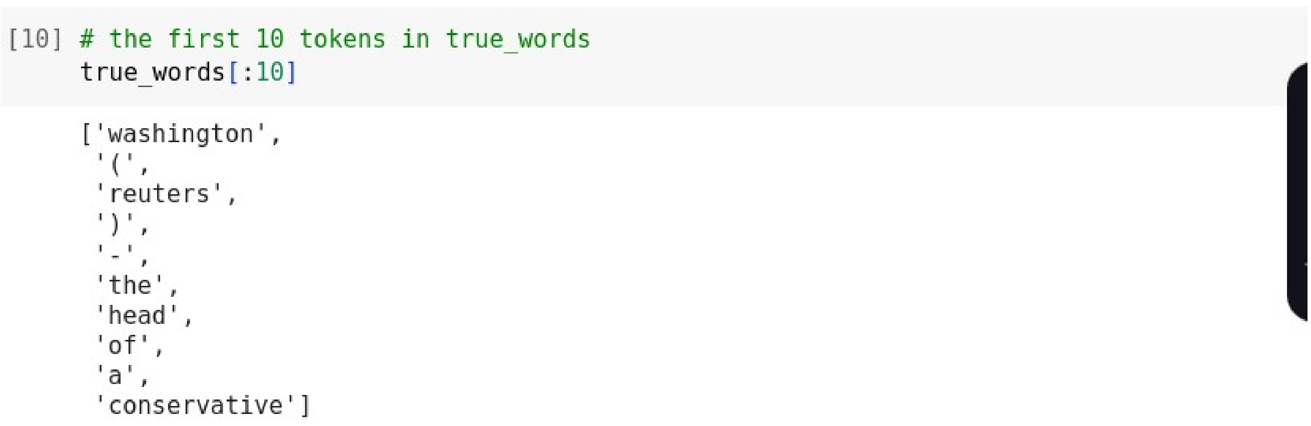
2. Lowercase

To guarantee that words are treated consistently regardless of case, all text has been converted to lowercase. In many NLP tasks, for instance, "Apple" and "apple" should be treated as the same word.
Lowercase text is more consistent and has less dimension in later processing steps because to this conversion.



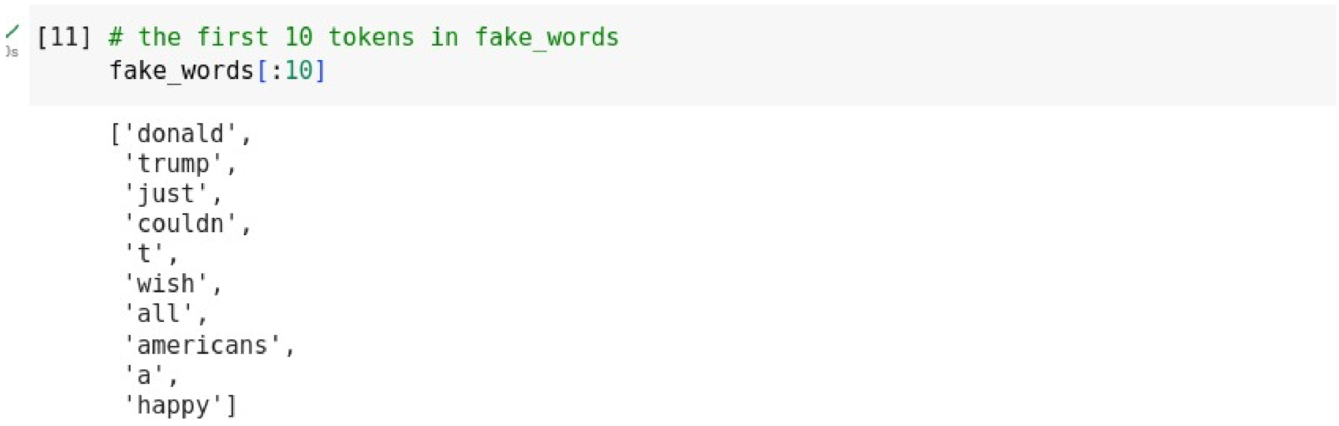
```
[7] # Next, we will use a list comprehension to process each token into lowercase.
    true_words = [w.lower() for w in true_tokens]
    fake_words = [w.lower() for w in fake_tokens]

    print(len(true_words))
    print(len(fake_words))
```

```
9320102
11036979
```



```
[10] # the first 10 tokens in true_words
     true_words[:10]
```

```
['washington',
 '(',
 'reuters',
 ')',
 '-',
 'the',
 'head',
 'of',
 'a',
 'conservative']
```

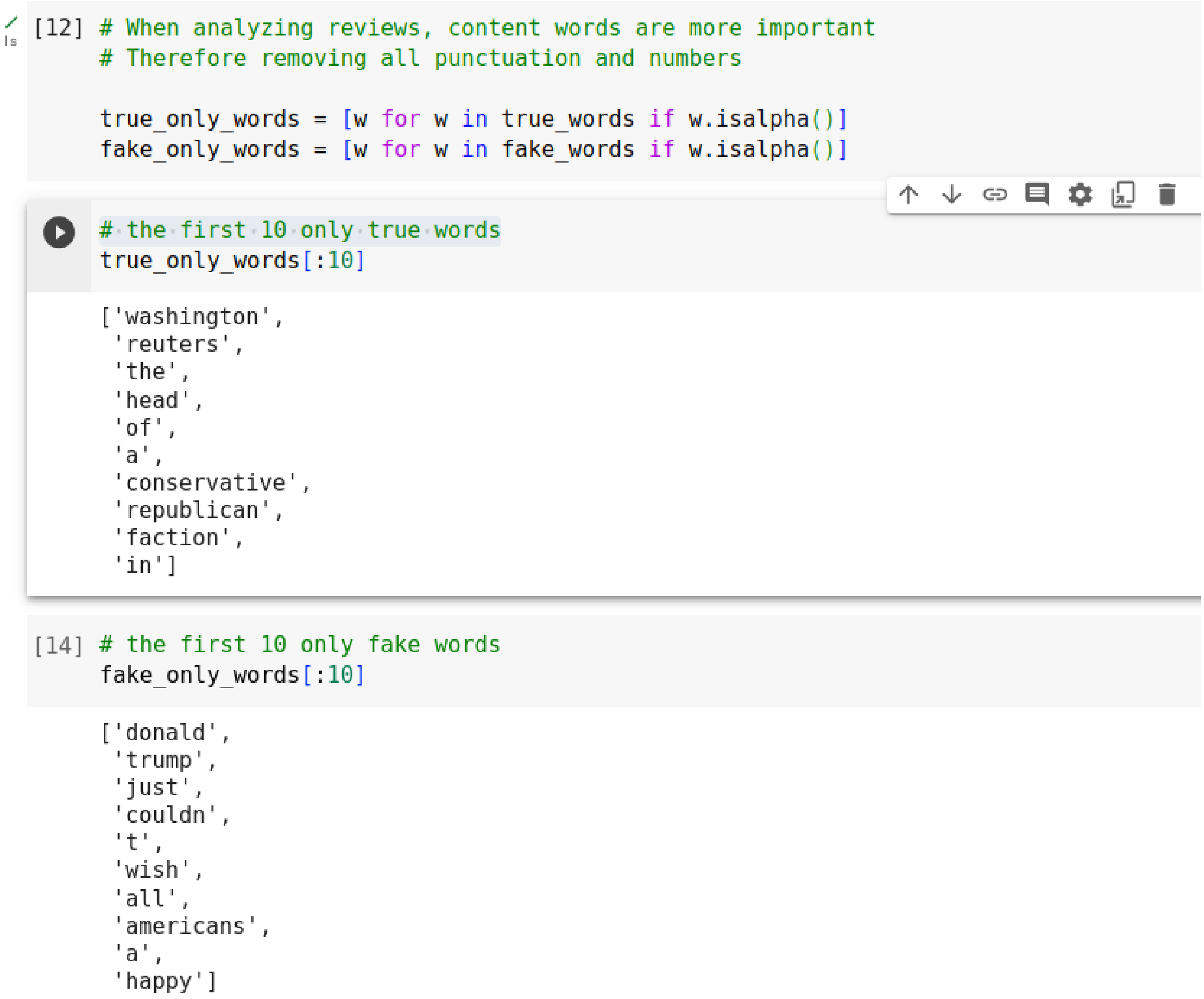


```
[11] # the first 10 tokens in fake_words
     fake_words[:10]
```

```
['donald',
 'trump',
 'just',
 'couldn',
 't',
 'wish',
 'all',
 'americans',
 'a',
 'happy']
```

3. Removing punctuation and numbers

In many NLP tasks, punctuation and numerals frequently don't convey any meaning. Eliminating them can aid in lowering text data noise.
By focusing on important terms, this stage simplifies the data and can increase text analysis's accuracy.



```
[12] # When analyzing reviews, content words are more important
     # Therefore removing all punctuation and numbers

     true_only_words = [w for w in true_words if w.isalpha()]
     fake_only_words = [w for w in fake_words if w.isalpha()]
```

```
# the first 10 only true words
true_only_words[:10]
```

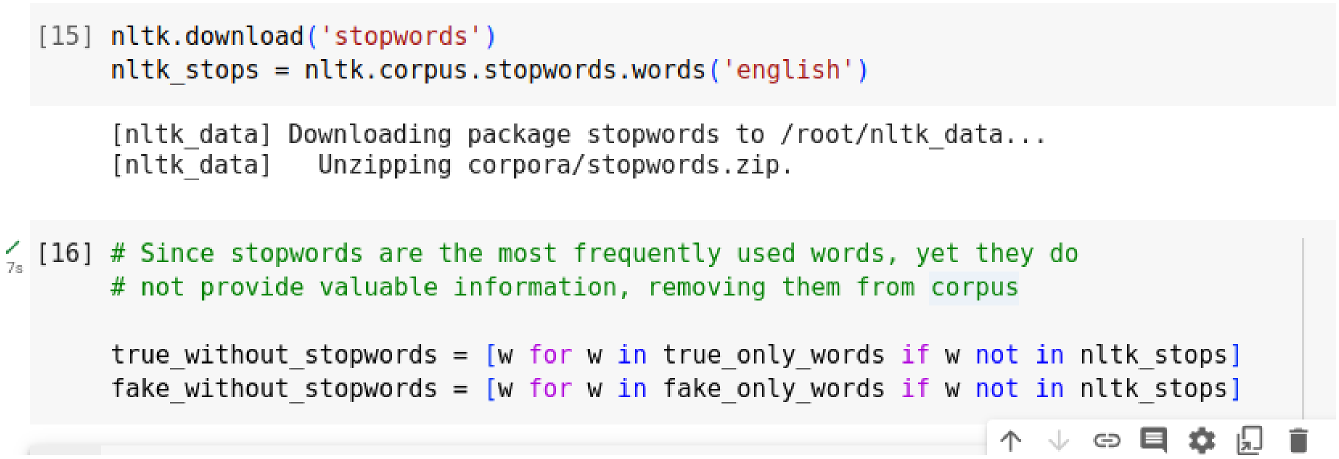
```
['washington',
 'reuters',
 'the',
 'head',
 'of',
 'a',
 'conservative',
 'republican',
 'faction',
 'in']
```

```
[14] # the first 10 only fake words
     fake_only_words[:10]
```

```
['donald',
 'trump',
 'just',
 'couldn',
 't',
 'wish',
 'all',
 'americans',
 'a',
 'happy']
```

4. Removing stopwords

Stopwords are frequent words (such as "the," "is," and "and") that are not necessary for many NLP tasks. Eliminating them can decrease dimensionality and boost analysis precision.
It may be difficult to distinguish between true and fake news when stopwords dominate word frequency counts.

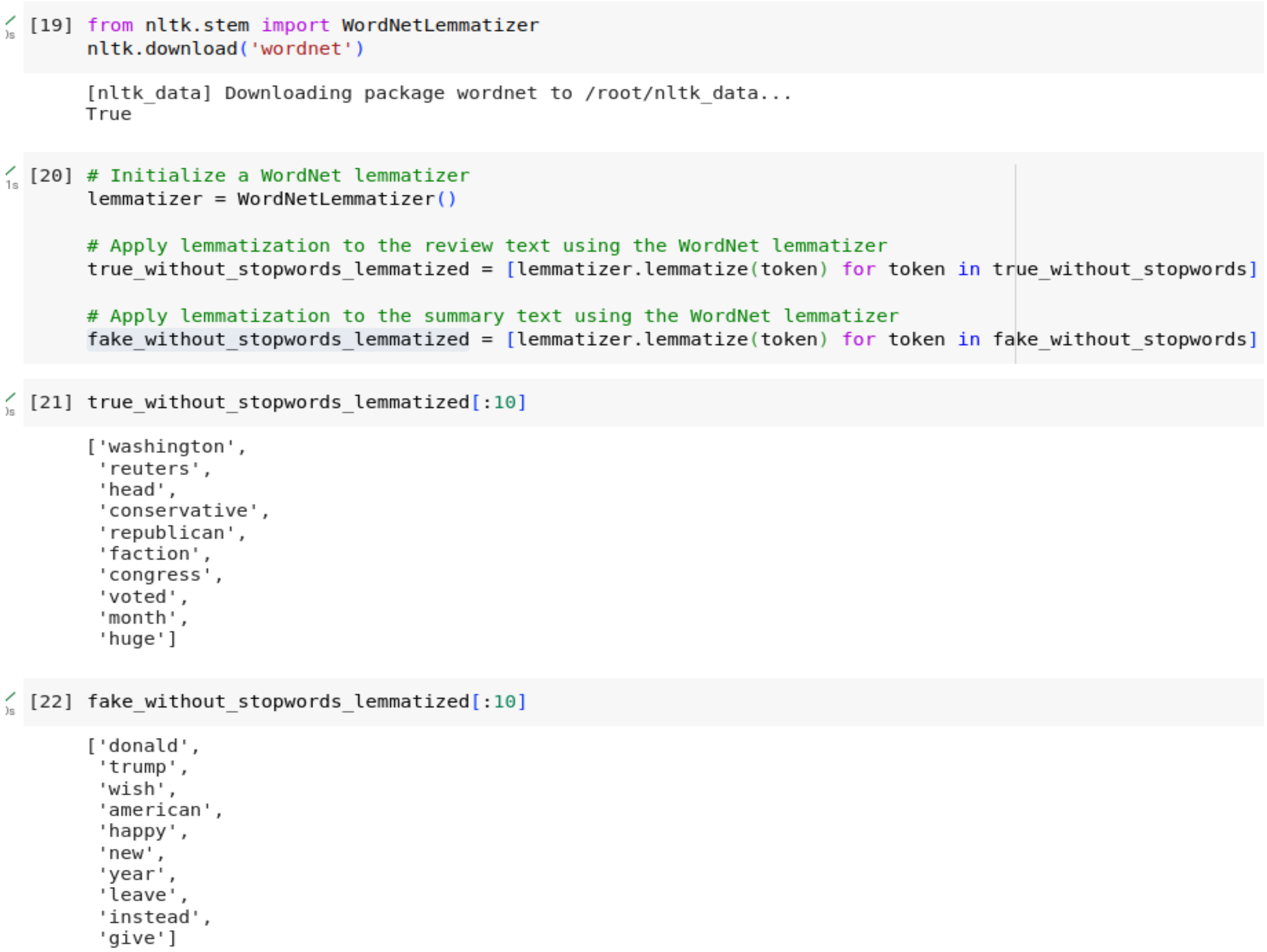


```
[15] nltk.download('stopwords')
     nltk_stops = nltk.corpus.stopwords.words('english')

     [nltk_data] Downloading package stopwords to /root/nltk_data...
     [nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[16] # Since stopwords are the most frequently used words, yet they do
     # not provide valuable information, removing them from corpus

     true_without_stopwords = [w for w in true_only_words if w not in nltk_stops]
     fake_without_stopwords = [w for w in fake_only_words if w not in nltk_stops]
```

5. Lemmatization

Words are reduced to their dictionary or basic form through lemmatization. This makes it easier to treat word variations like "running" and "ran" as the same word.
It makes sure that words are portrayed consistently, which makes it simpler to spot word links and patterns.



```
[19] from nltk.stem import WordNetLemmatizer
     nltk.download('wordnet')

     [nltk_data] Downloading package wordnet to /root/nltk_data...
     True
```

```
[20] # Initialize a WordNet lemmatizer
     lemmatizer = WordNetLemmatizer()

     # Apply lemmatization to the review text using the WordNet lemmatizer
     true_without_stopwords_lemmatized = [lemmatizer.lemmatize(token) for token in true_without_stopwords]

     # Apply lemmatization to the summary text using the WordNet lemmatizer
     fake_without_stopwords_lemmatized = [lemmatizer.lemmatize(token) for token in fake_without_stopwords]
```

```
[21] true_without_stopwords_lemmatized[:10]

     ['washington',
      'reuters',
      'head',
      'conservative',
      'republican',
      'faction',
      'congress',
      'voted',
      'month',
      'huge']
```

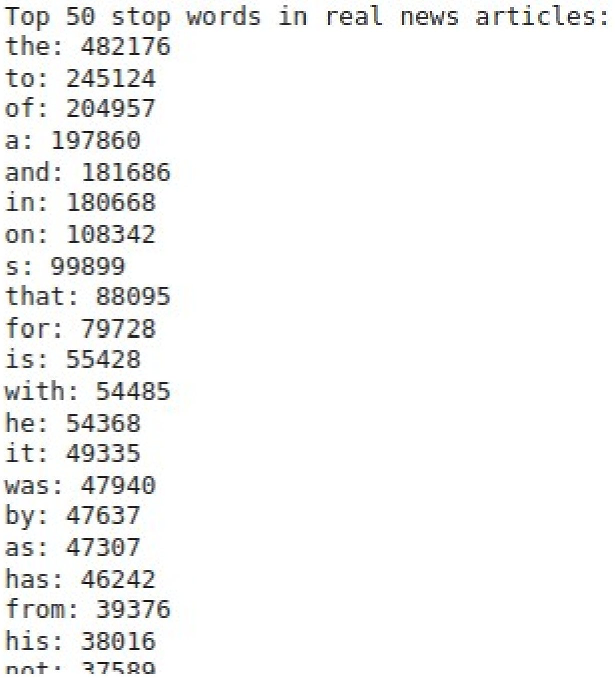
```
[22] fake_without_stopwords_lemmatized[:10]

     ['donald',
      'trump',
      'wish',
      'american',
      'happy',
      'new',
      'year',
      'leave',
      'instead',
      'give']
```

Data Analysis

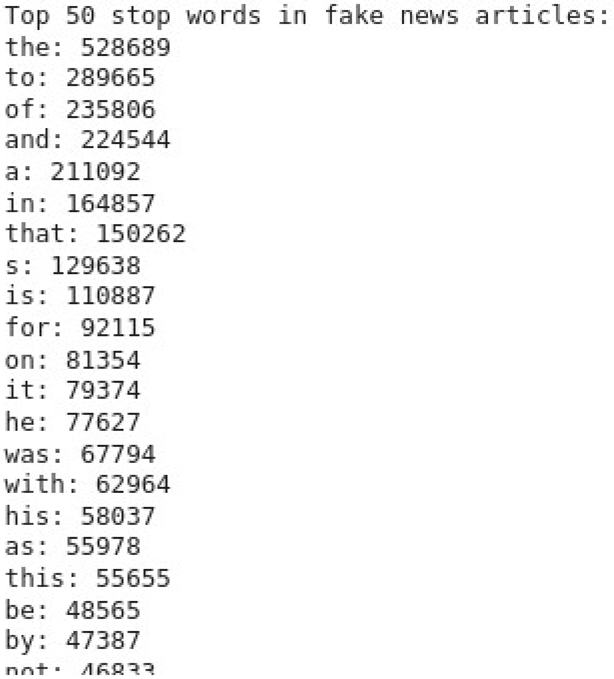
A key idea in Natural Language Processing (NLP) is word frequency, which describes how frequently a term appears in a corpus of texts. This metric is an important gauge of a word's importance inside the text in which it appears. Practically speaking, a word's frequency inside a document usually denotes its importance to the meaning of the document. These high-frequency words thus tend to be given more weight in the feature representation of the text. On the other hand, words with sporadic occurrences are generally thought to be less important and, in some situations, are classified as stop words to be omitted from the study.

1- List the top 50 stop words by frequency in fake news articles and those in real news articles.

The goal of this task is to list the top 50 stop words in both fake and true news articles. A list of the top 50 stop words for each category, together with the stop word's frequency of use, is provided by the code's output. Understanding the linguistic differences between the two categories of news stories will help with future analysis, such as text categorization or sentiment analysis.
You may compare the patterns of stop word usage between the two groups by listing the top stop words separately for fake and actual news stories. This comparison may highlight variations in writing or substance that could be a sign of fake versus true news.

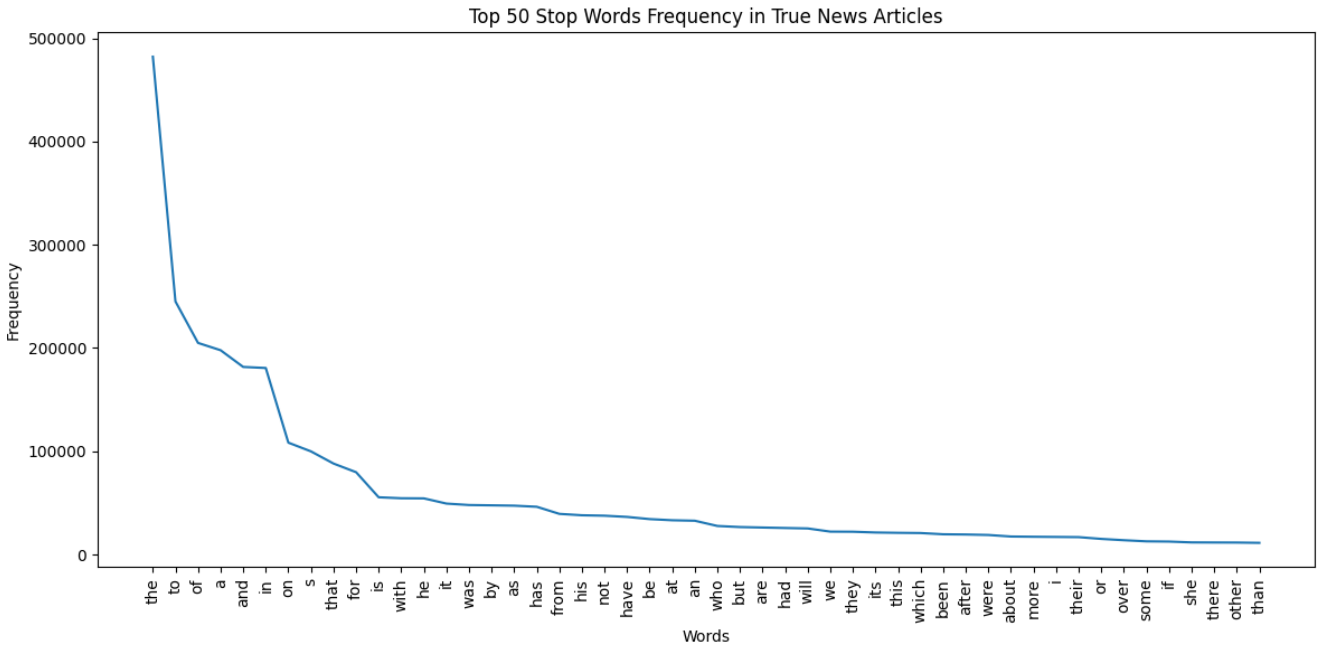


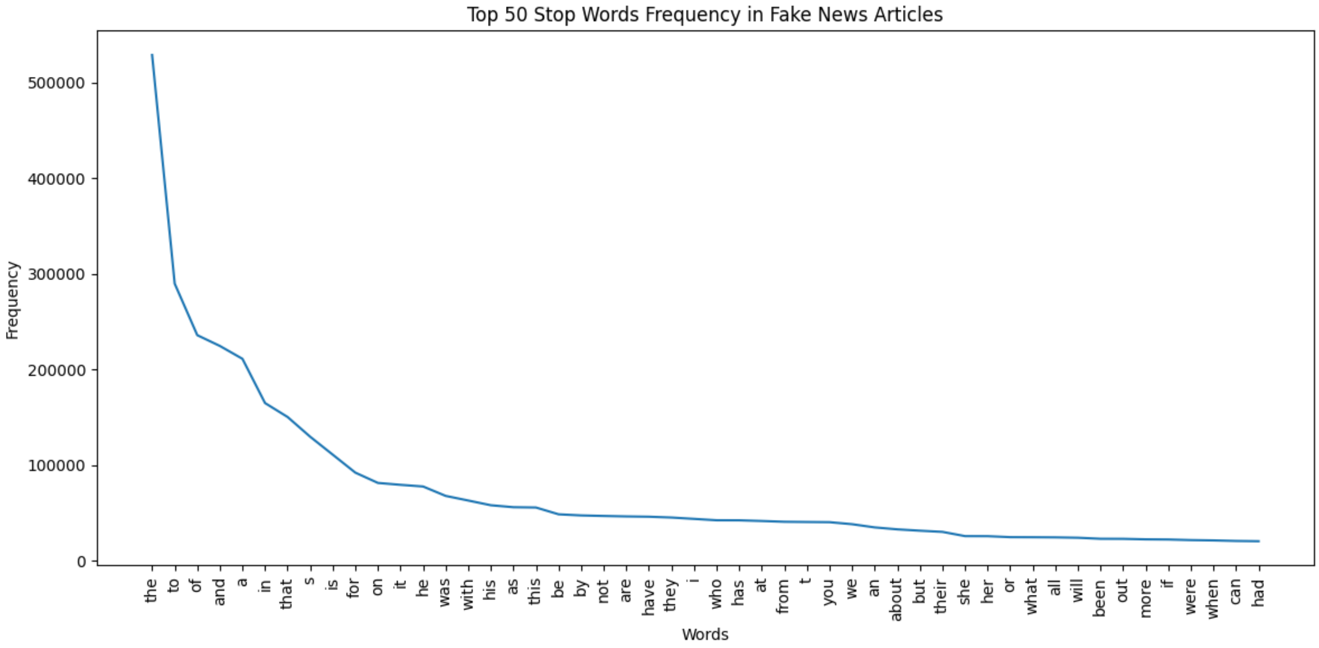
```
Top 50 stop words in real news articles:
the: 482176
to: 245124
of: 204957
a: 197860
and: 181686
in: 180668
on: 108342
s: 99899
that: 88095
for: 79728
is: 55428
with: 54485
he: 54368
it: 49335
was: 47940
by: 47637
as: 47307
has: 46242
from: 39376
his: 38016
not: 37589
```



```
Top 50 stop words in fake news articles:
the: 528689
to: 289665
of: 235806
and: 224544
a: 211092
in: 164857
that: 150262
s: 129638
is: 110887
for: 92115
on: 81354
it: 79374
he: 77627
was: 67794
with: 62964
his: 58037
as: 55978
this: 55655
be: 48565
by: 47387
not: 46833
```

Below is also the graphical presentation of the top 50 stop words frequency in both true and fake articles.





In both lists we can see that stop words such as “the”, “to”, “a”, “of”, “and” are the 5 most repeated.
We can see just a small difference in the graphs where the stop words on the 6 and 7 positions in true articles are used more than the stop words in the 6 and 7 position in fake articles.

2- List the top 50 content words by frequency in fake news articles and those in real news articles.

When we want to rank the top 50 content words by frequency in both fake news articles and real news articles, lemmatization is often preferred to stemming.

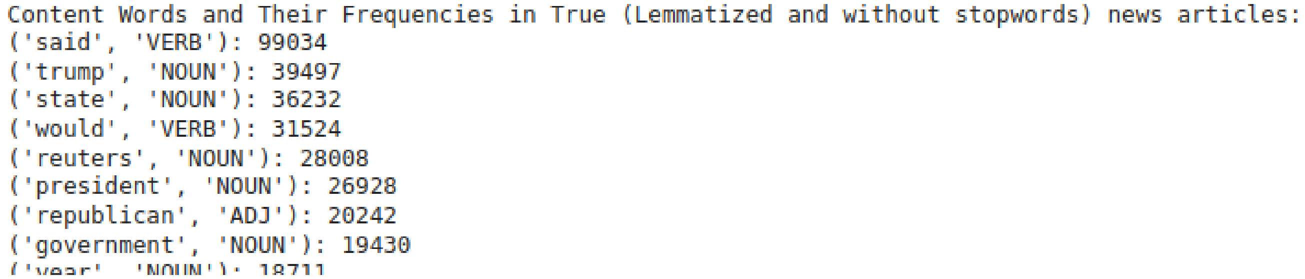
A term is "lemmatized" when it is reduced to its most basic or dictionary form, or "lemma." For instance, it would transform "running" to "run," "better" to "good," and "mice" to "mouse." This approach is more linguistically sophisticated and seeks to restore words to their semantic root form. It is very useful when trying to understand the terms' actual meanings.

To discover the word's underlying form, stemming, on the other hand, includes eliminating prefixes or suffixes from words. It can be more forceful than lemmatization and may not necessarily result in actual words. For example, it might transform "better" to "better" and "jumps" to "jump." Occasionally, stemming might produce terminology that are incorrect or nonstandard.

In the context of our question, we are interested in the frequency of content terms (likely nouns, verbs, adjectives, and adverbs) in news articles. It is advised to keep these terms in their dictionary or basic form rather than stemming them because they are more significant and useful for comprehending the content.

The result, which is based on the lemmatized and preprocessed text data without stopwords, lists the top 50 content words by frequency in both false and real news items.

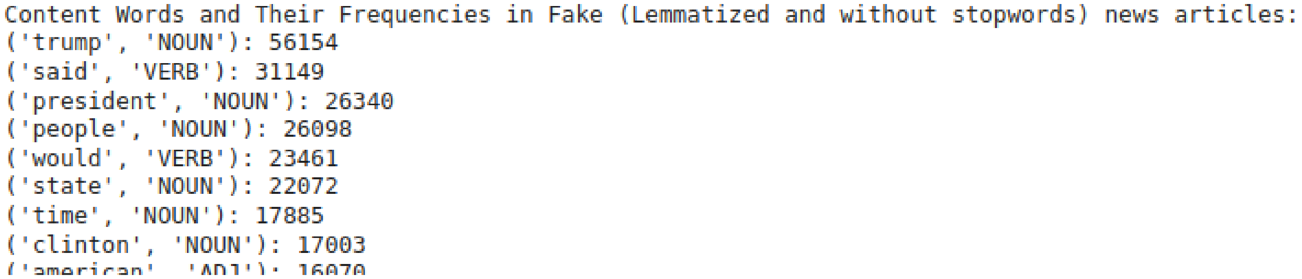
In articles with real news:



```
Content Words and Their Frequencies in True (Lemmatized and without stopwords) news articles:
('said', 'VERB'): 99034
('trump', 'NOUN'): 39497
('state', 'NOUN'): 36232
('would', 'VERB'): 31524
('reuters', 'NOUN'): 28008
('president', 'NOUN'): 26928
('republican', 'ADJ'): 20242
('government', 'NOUN'): 19430
('year', 'NOUN'): 18711
```

The most frequently used content word is "said," which is followed by "trump," "state," and "would."
These terms, which include remarks (said), the president of the United States (trump), official government activities (state), and hypothetical situations (would), are frequently used in news reporting.
Other words with a high frequency of usage in content include "president," "republican," "government," "year," "house," "new," and "people."
As might be expected in authentic news items, many of these terms have to do with political and governmental issues.

In articles with fake news:



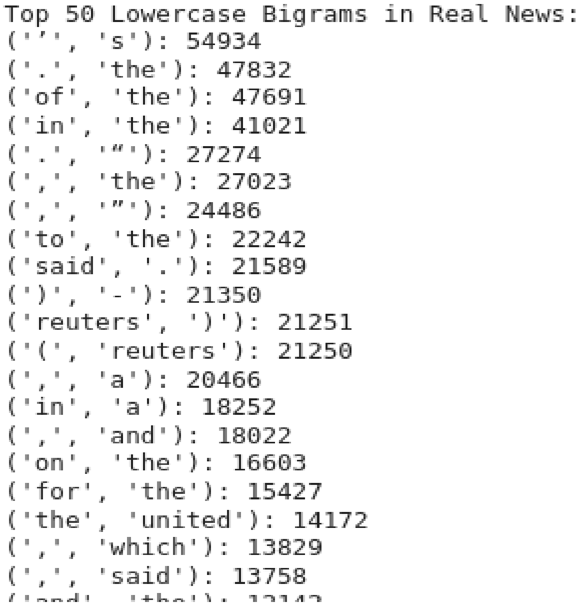
```
Content Words and Their Frequencies in Fake (Lemmatized and without stopwords) news articles:
('trump', 'NOUN'): 56154
('said', 'VERB'): 31149
('president', 'NOUN'): 26340
('people', 'NOUN'): 26098
('would', 'VERB'): 23461
('state', 'NOUN'): 22072
('time', 'NOUN'): 17885
('clinton', 'NOUN'): 17003
('american', 'ADJ'): 16070
```

Trump is the term that appears in content the most, followed by "said," "president," and "people."
A emphasis on the U.S. President is shown by the fact that "trump" is the word used most frequently in the content of fake news items.
Among the other frequently used words are "state," "time," "Clinton," "American," "also," and "say."
The president of the United States, political figures (such as Hillary Clinton), and broad comments (say) appear to be highlighted more frequently in fake news pieces.

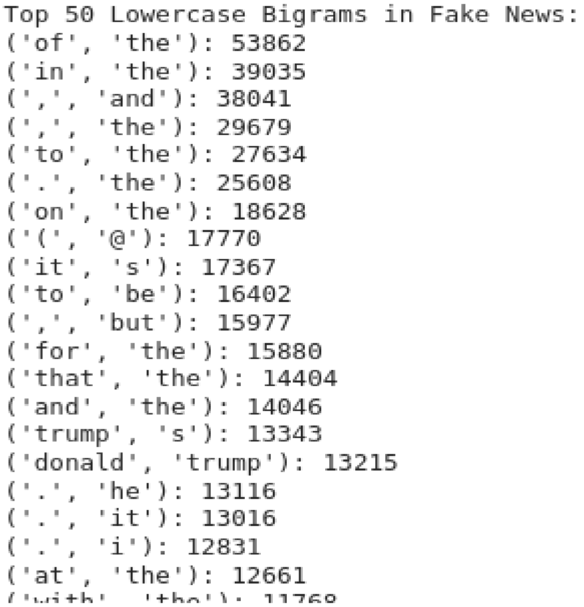
Overall, the topics and themes prevalent in news reporting are reflected in the content words of both real and fraudulent news stories. While fake news articles seem to place a strong emphasis on certain personalities and political figures, particularly the U.S. President ("trump"), real news pieces cover a wider range of political and governmental themes.

3- List the top 50 bigrams by frequencies in fake news articles and those in real news articles.

Real or fake news reports can be examined for bigrams (two adjacent words), which can reveal important grammatical and structural distinctions between the two types of content.
We may find frequently used terms in both true and fraudulent news by looking at the top bigrams. Differences in writing style and tone can be revealed through bigrams.



```
Top 50 Lowercase Bigrams in Real News:
('’', 's'): 54934
('.', 'the'): 47832
('of', 'the'): 47691
('in', 'the'): 41021
('.', '“'): 27274
(',', 'the'): 27023
(',', '”'): 24486
('to', 'the'): 22242
('said', '.'): 21589
(')', '-'): 21350
('reuters', ')'): 21251
('(', 'reuters'): 21250
(',', 'a'): 20466
('in', 'a'): 18252
(',', 'and'): 18022
('on', 'the'): 16603
('for', 'the'): 15427
('the', 'united'): 14172
(',', 'which'): 13829
(',', 'said'): 13758
('and', 'the'): 13143
```

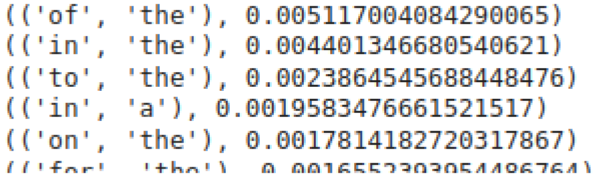


```
Top 50 Lowercase Bigrams in Fake News:
('of', 'the'): 53862
('in', 'the'): 39035
(',', 'and'): 38041
(',', 'the'): 29679
('to', 'the'): 27634
('.', 'the'): 25608
('on', 'the'): 18628
('(', '@'): 17770
('it', 's'): 17367
('to', 'be'): 16402
(',', 'but'): 15977
('for', 'the'): 15880
('that', 'the'): 14404
('and', 'the'): 14046
('trump', 's'): 13343
('donald', 'trump'): 13215
('.', 'he'): 13116
('.', 'it'): 13016
('.', 'i'): 12831
('at', 'the'): 12661
('with', 'the'): 11768
```

The bigrams above are without removing the stop words so most of them do not make sense at all.
The next step is to filter the bigrams that include the punctuation marks or other non-alphabetic tokens and filter those that include the stopwords.

True News:

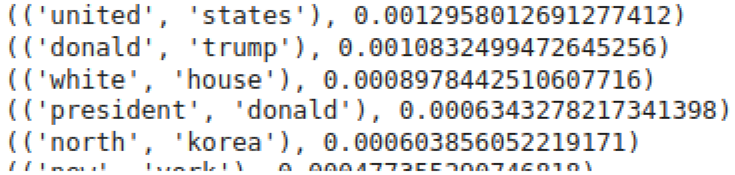
The initial section of the code figures out how frequently bigrams appear in true news. In other words, it keeps track of how frequently each distinct bigram appears in the text. Here are the top five bigrams in this phase and their interpretations:



```
(('of', 'the'), 0.005117004084290065)
(('in', 'the'), 0.004401346680540621)
(('to', 'the'), 0.0023864545688448476)
(('in', 'a'), 0.0019583476661521517)
(('on', 'the'), 0.0017814182720317867)
(('for', 'the'), 0.0016552393954486764)
```

('of', 'the'): This bigram shows the frequent use of the preposition "of" followed by the article "the" in real news articles, where it occurs about 0.51% of the time.
('in', 'the'): Similar to the previous bigram, the phrase "in the" occurs frequently, or 0.44% of the time.
('to', 'the'): In news articles, the word "to the" appears a lot; it makes up roughly 0.24% of bigrams.
('in', 'a'): The bigram "in a" appears in about 0.20 percent of the text, illustrating how frequently articles are used.
('on', 'the'): "On the" appears about 0.18% of the time, which is consistent with the frequency of expressions referring to events or occurrences.

The second part of the code ensures that only bigrams made up of alphabetic characters are taken into consideration by filtering out non-alphabetical bigrams. The top five filtered bigrams for true news are interpreted as follows:



```
(('united', 'states'), 0.0012958012691277412)
(('donald', 'trump'), 0.0010832499472645256)
(('white', 'house'), 0.0008978442510607716)
(('president', 'donald'), 0.0006343278217341398)
(('north', 'korea'), 0.0006038560522191171)
(('new', 'york'), 0.000477355200746818)
```

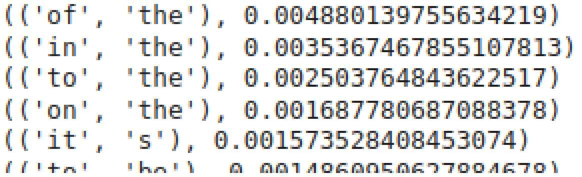
('united','states'): Even after filtering, the bigram "united states" still appears the most frequently, demonstrating its importance in actual news items.
('donald', 'trump'): The name "Donald Trump" appears frequently in actual news items, indicating the concentration on public officials.
('white,' 'house'): This bigram illustrates how significant the White House is in news stories.
('president', 'donald'): It is clear that the President is frequently mentioned in the news.

We ensure that significant word combinations are taken into account by removing non-alphabetical bigrams, which aids in recognizing important themes and figures.

The results that appear include bigrams like "new york," "prime minister," "told reuters," and "u.s. president," which offer details on a variety of subjects like locales, public people, and news organizations.

Fake News:

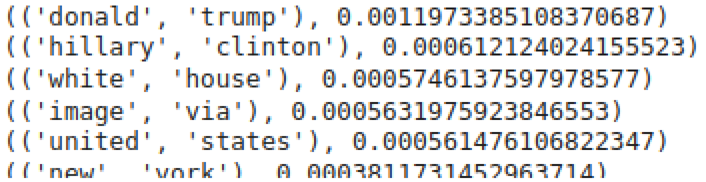
The top five bigrams in this phase and their interpretations:



```
(('of', 'the'), 0.004880139755634219)
(('in', 'the'), 0.0035367467855107813)
(('to', 'the'), 0.002503764843622517)
(('on', 'the'), 0.001687780687088378)
(('it', 's'), 0.001573528408453074)
(('to', 'be'), 0.0014860950627884678)
```

('of', 'the'): This bigram shows how frequently the preposition "of" is followed by the article "the," appearing in bogus news articles about 0.49 percent of the time.
('in', 'the'): Similar to real news, the phrase "in the" occurs frequently, roughly 0.35% of the time.
('to', 'the'): In around 0.25% of bigrams in bogus news items, the word "to the" appears frequently.
('on', 'the'): The phrase "on the" appears in about 0.17% of the text, which reflects the frequency of expressions referring to events or occurrences.
('it', 's'): Its frequent usage makes the contraction "it's" one of the most common bigrams in fake news.

In the second part of the code, I have removed filtered bigrams without stop words. The top five bigrams that were stop word filtered are explained as follows:



```
(('donald', 'trump'), 0.0011973385108370687)
(('hillary', 'clinton'), 0.000612124024155523)
(('white', 'house'), 0.0005746137597978577)
(('image', 'via'), 0.0005631975923846553)
(('united', 'states'), 0.000561476106822347)
(('new', 'york'), 0.0003811731452963714)
```

('donald', 'trump'): The name "Donald Trump" is still very common, just like in the unfiltered results.
('hillary,' 'clinton') The name "Hillary Clinton" continues to be used often.
('white,' 'house') The significance of the White House is still present in fake news.
('image,' 'via') Fake news continues to cite the origin of the photographs.
('united,''states'): The phrase "United States" is frequently used.

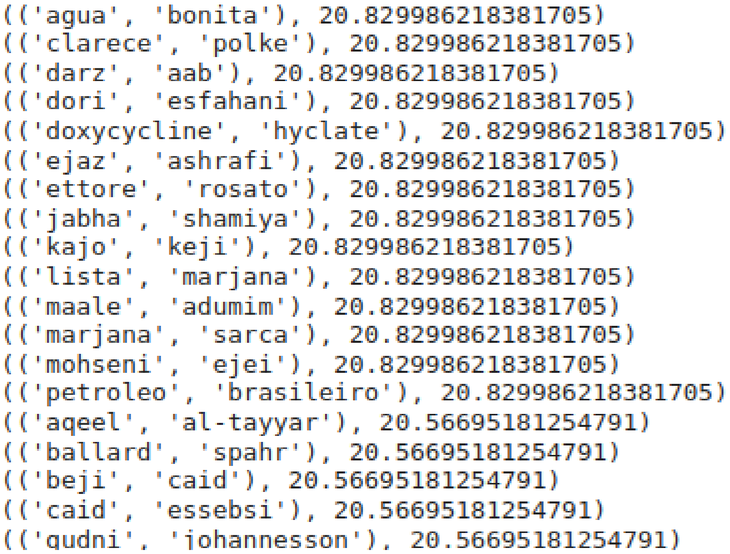
The analysis narrows its emphasis to bigrams with more significant substance by removing stop words.

Bigrams like "new york," "president obama," "fox news," "climate change," "health care," and more can be found in the results that appear after. These bigrams reflect a variety of issues, including places, political leaders, news organizations, and niche topics like health care and climate change.

4- List the top 50 bigrams by their Mutual Information scores (using min frequency 5) in fake news articles and those in real news articles

This task lists the top 50 bigrams by their Mutual Information (MI) scores for both fake news articles and real news articles. MI is a measure of association between two words, and it can help identify significant word pairs that co-occur more frequently than expected by chance.

Top PMI Scores for True News Articles:
The top 50 bigrams in actual news items with the highest MI ratings are as follows:



```
(('agua', 'bonita'), 20.829986218381705)
(('clarece', 'polke'), 20.829986218381705)
(('darz', 'aab'), 20.829986218381705)
(('dori', 'esfahani'), 20.829986218381705)
(('doxycycline', 'hyclate'), 20.829986218381705)
(('ejaz', 'ashrafi'), 20.829986218381705)
(('ettore', 'rosato'), 20.829986218381705)
(('jabha', 'shamiya'), 20.829986218381705)
(('kajo', 'keji'), 20.829986218381705)
(('lista', 'marjana'), 20.829986218381705)
(('maale', 'adumim'), 20.829986218381705)
(('marjana', 'sarca'), 20.829986218381705)
(('mohseni', 'ejei'), 20.829986218381705)
(('petroleo', 'brasileiro'), 20.829986218381705)
(('aqeel', 'al-tayyar'), 20.56695181254791)
(('ballard', 'spahr'), 20.56695181254791)
(('beji', 'caid'), 20.56695181254791)
(('caid', 'essebsi'), 20.56695181254791)
(('gudni', 'johannesson'), 20.56695181254791)
```

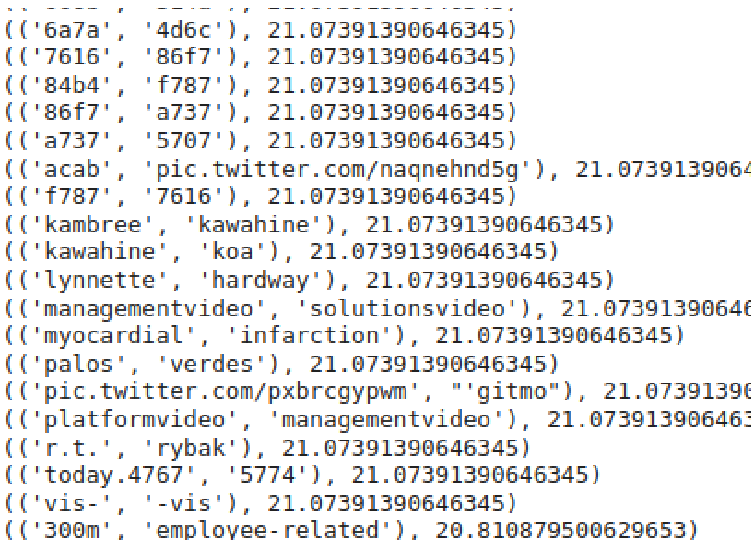
('agua', 'bonita'), 'clarece', 'polke', 'darz', 'aab', and additional words like these: The exceptionally high MI scores for these bigrams indicate a highly potent relationship. They might, however, be domain-specific or uncommon words, and their high MI may be a result of their infrequent usage in the text.

('marjana','sarca'), ('mohseni', 'ejei'), ('moqtada', 'al-sadr'): These bigrams are connected to names of people or particular words. MI emphasizes their importance in the context of actual news reports.

('hillary,' 'clinton') The name "Hillary Clinton" has a high MI score in actual news items, which is to be expected and shows its strong linkage in this context.

('supreme', 'court'), ('foreign','ministry'), and ('human', 'rights') These bigrams connect to important and frequently discussed subjects in the actual news.

Top PMI Scores for Fake News Articles:
The top 50 bigrams in false news stories with the highest MI scores are as follows:



```
(('6a7a', '4d6c'), 21.07391390646345)
```

```
(('7616', '86f7'), 21.07391390646345)
(('84b4', 'f787'), 21.07391390646345)
(('86f7', 'a737'), 21.07391390646345)
(('a737', '5707'), 21.07391390646345)
(('acab', 'pic.twitter.com/naqnehnd5g'), 21.0739139064
(('f787', '7616'), 21.07391390646345)
(('kambree', 'kawahine'), 21.07391390646345)
(('kawahine', 'koa'), 21.07391390646345)
(('lynnette', 'hardway'), 21.07391390646345)
(('managementvideo', 'solutionsvideo'), 21.07391390646
(('myocardial', 'infarction'), 21.07391390646345)
(('palos', 'verdes'), 21.07391390646345)
(('pic.twitter.com/pxbrcgypwm', "'gitmo"), 21.07391390
(('platformvideo', 'managementvideo'), 21.073913906463
(('r.t.', 'rybak'), 21.07391390646345)
(('today.4767', '5774'), 21.07391390646345)
(('vis-', '-vis'), 21.07391390646345)
(('300m', 'employee-related'), 20.810879500629653)
```

('0000', '0907'), ('//t.co/ltdtbehhgh', 'pic.twitter.com/t2s8ufif5o'), and others: Despite having incredibly high MI scores, these bigrams appear to be incomprehensible or gibberish. They might be data artifacts or artificially created content.

('managementvideo,''solutionsvideo'): This bigram has a high MI score, but without more context, its meaning is ambiguous.

('mata', 'pires'), ('terrell','mcsweeny'), and ('vis-', '-vis') Although some names and terms are included in these bigrams, they might not be well known. Because they are uncommon in bogus news stories, the high MI scores might be explained.

('fern', 'ndez'), ('yik', 'yak'), ('bucolic', 'adirondacks') Although these bigrams are connected to certain words and places, they might not be widely used or especially significant in the context of fake news.

('paolo', 'gentiloni'): Although "Paolo Gentiloni" has a high MI score, its relevance may change based on the false information's contents.

Overall, MI scores assist in identifying word associations in the context of legitimate and false news. High MI scores are sometimes obtained in false news for possibly absurd or uncommon phrases, whereas in real news, high MI scores are typically obtained for words and terms that are significant and meaningful.

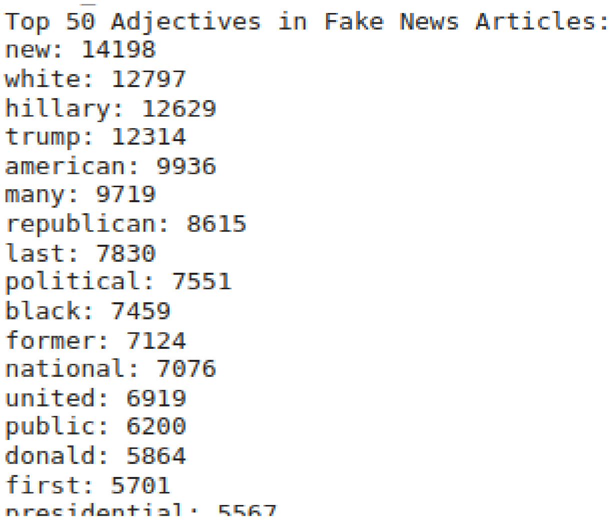
5- List top 50 adjective words in fake news articles and those in real news articles.

The given code examines the most prevalent adjectives found in both fake and true news items. Adjectives are detailed words that give nouns additional context.

Fake News

The data shows that adjectives in false news articles are typically more general, intense, and emotive, and they commonly focus on significant personalities like Hillary

Clinton and Donald Trump. These usually arouse emotions while giving the story a sense of urgency.

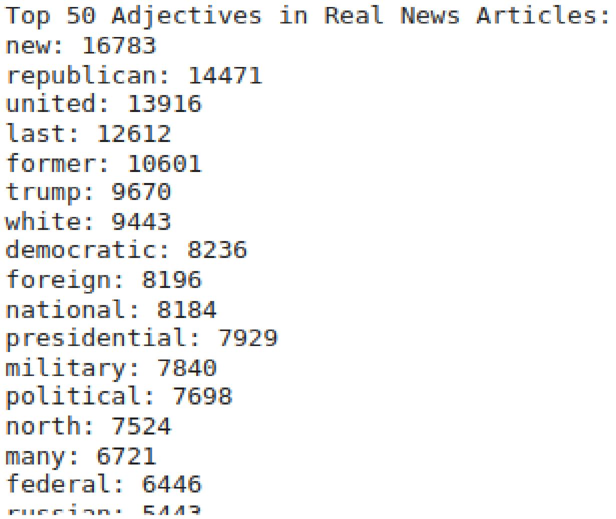


```
Top 50 Adjectives in Fake News Articles:
new: 14198
white: 12797
hillary: 12629
trump: 12314
american: 9936
many: 9719
republican: 8615
last: 7830
political: 7551
black: 7459
former: 7124
national: 7076
united: 6919
public: 6200
donald: 5864
first: 5701
presidential: 5567
```

Fake news frequently presents content as "new" to attract readers' interest and create a feeling of urgency.
White: This could be a reference to politics or matters pertaining to the White House.
hillary: The word "hillary" is often used to refer to rumors or disagreements regarding Hillary Clinton.
trump: In political fake news, the word "trump" is frequently used to suggest that the attention is on former President Donald Trump.
american: The word "american" may be used to express national concerns or invoke feelings of patriotism.

True News:

Real news pieces appear to use more diverse that are tailored to the news's context. They use words like "republican," "united," "presidential," "military," "foreign," and "senate," which are common in news stories on politics and world affairs. The words "nuclear," "economic," "legal," and "syrian" are present because more specific and important news subjects are being prioritized.

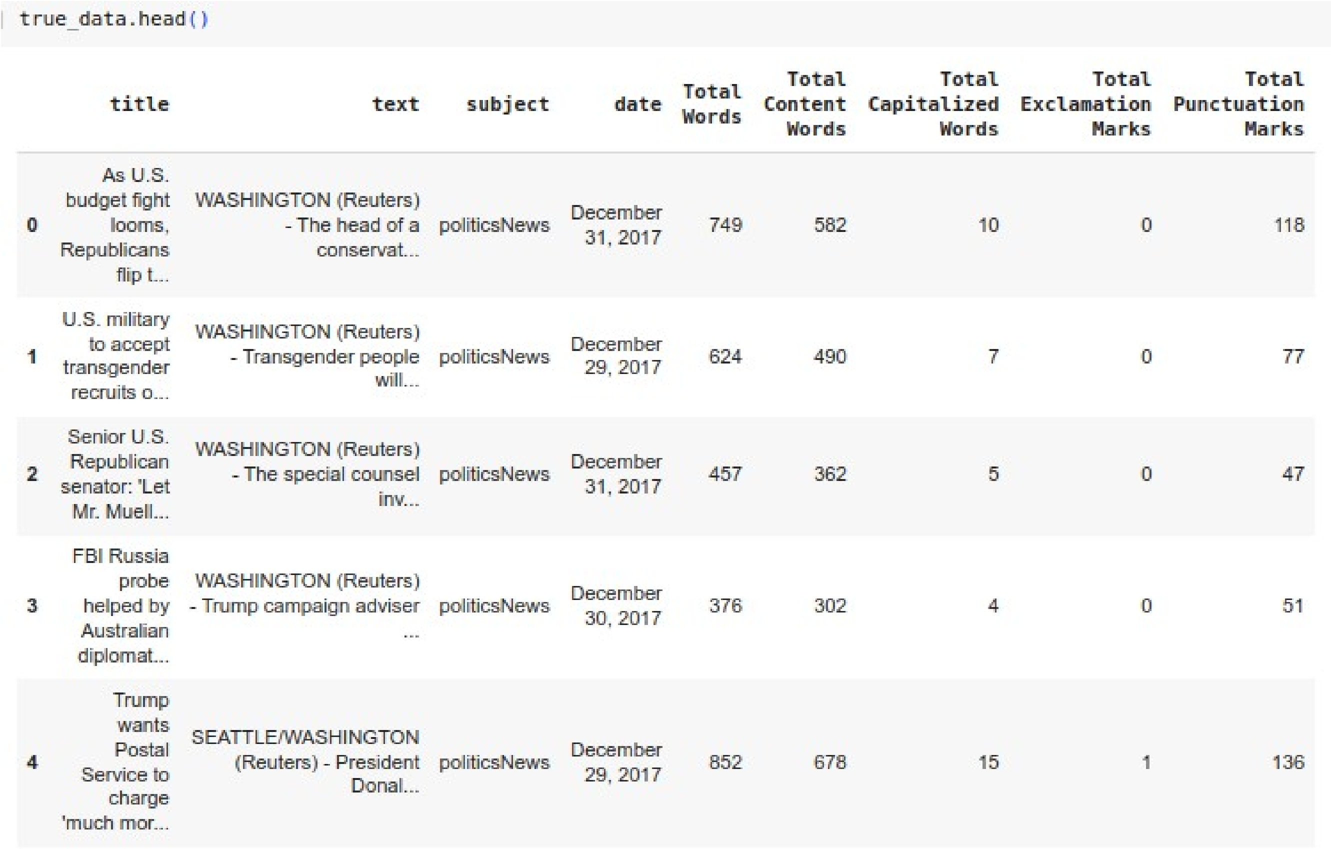


```
Top 50 Adjectives in Real News Articles:
new: 16783
republican: 14471
united: 13916
last: 12612
former: 10601
trump: 9670
white: 9443
democratic: 8236
foreign: 8196
national: 8184
presidential: 7929
military: 7840
political: 7698
north: 7524
many: 6721
federal: 6446
russian: 5443
```

In genuine news, the word "new" is frequently used to describe recent developments or happenings.
republican: The word "republican" is frequently used to indicate political prejudice or a focus on Republican-related issues.
presidential: This phrase applies to news stories that cover the presidency.
military: The word "military" is frequently used, which signals that the true news is on defense and the armed forces.
foreign: The word "foreign" is frequently used in genuine news articles and may refer to global issues.

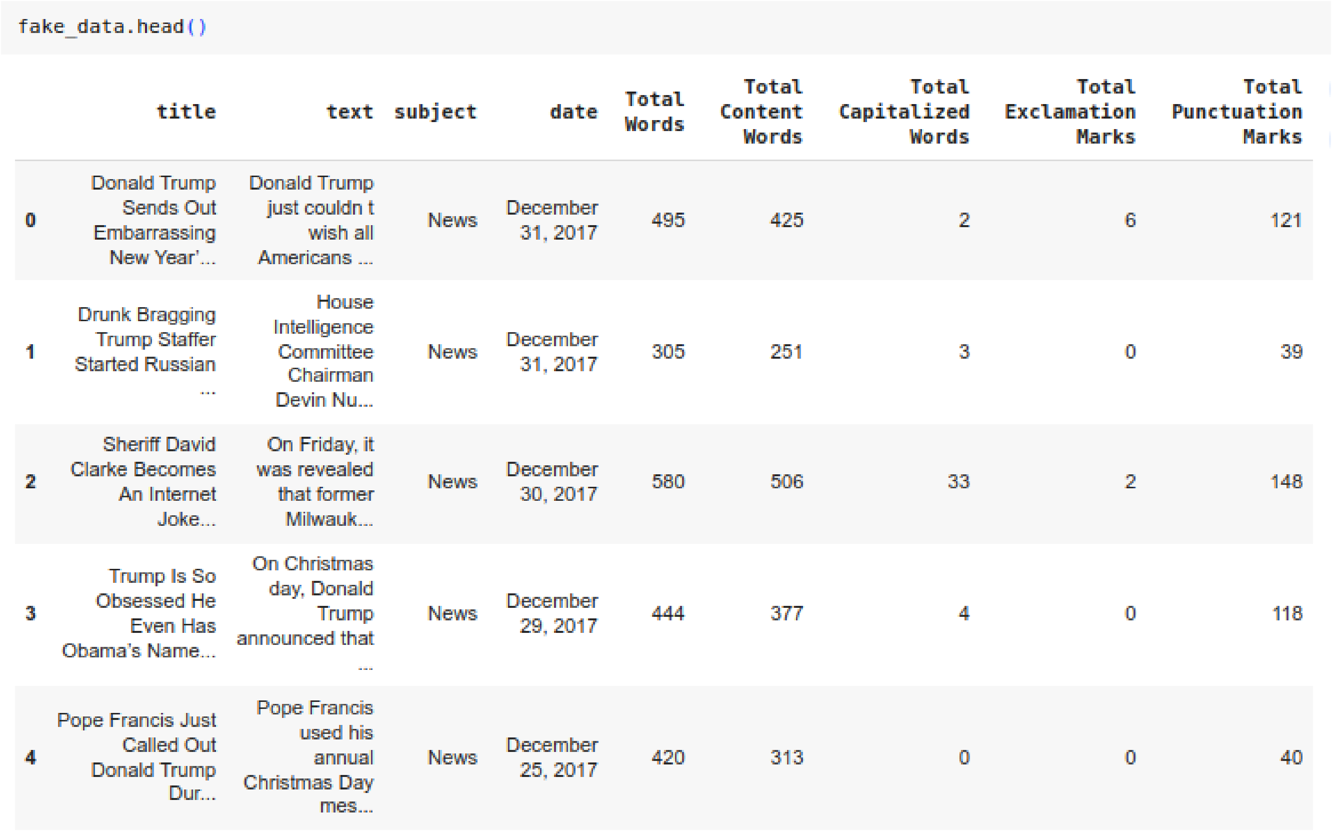
For each article, you will also obtain total number of words, total number of content words, total number of words that are all capitalized (excluding “I”), total number of exclamation marks, and total number of punctuation marks. You will save this information in CSV files by creating new columns and saving the information in these columns.

The two sets of articles saved in the true_data and fake_data DataFrames are processed as required. This step calculates and extracts data for each article, including the total amount of words, content words (apart from commonly used stop words), capitalized words (except from "I"), exclamation points, and punctuation. The revised DataFrames are exported to CSV files, enriching the original data for future analysis. This data is added as new columns to the DataFrames.



```
true_data.head()
```

	title	text	subject	date	Total Words	Total Content Words	Total Capitalized Words	Total Exclamation Marks	Total Punctuation Marks
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	749	582	10	0	118
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	624	490	7	0	77
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	457	362	5	0	47
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	376	302	4	0	51
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	852	678	15	1	136



```
fake_data.head()
```

	title	text	subject	date	Total Words	Total Content Words	Total Capitalized Words	Total Exclamation Marks	Total Punctuation Marks
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	495	425	2	6	121
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	305	251	3	0	39
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	580	506	33	2	148
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	444	377	4	0	118
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	420	313	0	0	40

Conclusion

In conclusion, this work on Python programming and corpus statistics in the context of Natural Language Processing (NLP) aims to solve the crucial problem of the dissemination of false information in the current digital era. It has gotten harder to tell fact from fiction with the ever-growing amount of information on the internet. Our capacity to distinguish between true and false news has been improved by using a variety of preprocessing techniques and data analysis approaches to analyze a dataset of both fake and real news pieces.

Data preparation and text analysis of the "ISOT Fake News Dataset," which consists of approximately 12,600 stories from reliable and questionable sources, were the main objectives of this work.

Tokenization, lowercase conversion, punctuation and number removal, stopword elimination, and lemmatization are a few of the essential stages that went into cleaning and preparing the data. These processes made sure that the text data was in an analysis-ready format.

From the text data, insights were to be drawn. The top 50 stop words in both fake and actual news items were ranked by frequency, illuminating the grammatical distinctions between the two groups. The top 50 content words by frequency were displayed, illuminating the recurring themes in news reporting. To comprehend the grammatical and structural differences between the two forms of material, bigrams were investigated. The top 50 bigrams with their MI scores were listed after meaningful word pairings were found using the Mutual Information (MI) score. The top 50 adjectives in both fake and legitimate news stories were also retrieved, giving information about the focus and tone of both groups.

The study went beyond standard text analysis by obtaining additional textual statistics for each article, such as the total number of words and the number of words in each paragraph. It also counted the number of capitalized words (aside from "I"), exclamation points, punctuation marks, and capitalized terms. These extra details improved the dataset for next research.

The findings of this study have shed important light on the differences between news stories that are real and those that are not. While there are certain differences between the two, such as fake news's emphasis on particular characters, both kinds of stories reflect general issues with news reporting.

The quality and dependability of the online information landscape can be improved by creating tools and tactics to recognize and combat false information by studying the linguistic and content trends in real and false news.