TDT4310 – LAB 1

Exercise 1

a) The words beginning with "sh" are she, shells, and shore. I found this by slicing off the two first characters of the word and checking whether or not it was "sh".

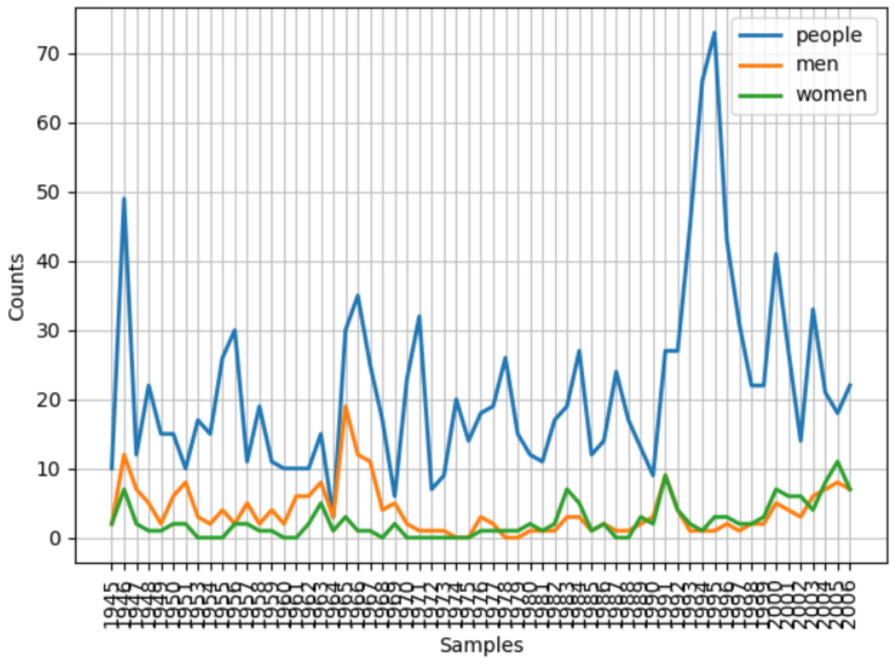
b) The words longer than four characters are sells, shells and shore. I found this by taking the length of each word and checking whether or not it was longer than four.

Exercise 2

a) I found the count by taking a conditional frequency distribution using the .txt files as experiments and the words as conditions, running each file in the corpus under the condition that the file contains the word I am looking for. For each experiment, I ran each file through each of the words, checking if any of the words in the file (converted to lower) are the same as the word I am looking for. I tabulated this with the tabulate() function, using the corpus files as conditions and the words as samples.

```
                        men  women people
    1945-Truman.txt       2      2     10
    1946-Truman.txt      12      7     49
    1947-Truman.txt       7      2     12
    1948-Truman.txt       4      1     22
    1949-Truman.txt       2      1     15
    1950-Truman.txt       6      2     15
    1951-Truman.txt       8      2      9
1953-Eisenhower.txt       3      0     17
1954-Eisenhower.txt       2      0     15
1955-Eisenhower.txt       4      0     26
1956-Eisenhower.txt       2      2     30
1957-Eisenhower.txt       5      2     11
1958-Eisenhower.txt       2      1     19
1959-Eisenhower.txt       4      1     11
1960-Eisenhower.txt       2      0     10
   1961-Kennedy.txt       6      0     10
   1962-Kennedy.txt       6      2     10
   1963-Johnson.txt       0      0      3
   1963-Kennedy.txt       8      5     12
   1964-Johnson.txt       3      1      3
 1965-Johnson-1.txt       7      0     16
 1965-Johnson-2.txt      11      3     14
 1966-Johnson.txt     12      1     35
 1967-Johnson.txt     11      1     25
 1968-Johnson.txt      4      0     17
 1969-Johnson.txt      5      2      6
   1970-Nixon.txt      2      0     23
   1971-Nixon.txt      1      0     31
   1972-Nixon.txt      1      0      7
   1973-Nixon.txt      0      0      9
   1974-Nixon.txt      0      0     19
    1975-Ford.txt      0      0     13
    1976-Ford.txt      3      1     18
    1977-Ford.txt      2      1     17
  1978-Carter.txt      0      1     26
  1979-Carter.txt      0      1     15
  1980-Carter.txt      1      2     11
  1981-Reagan.txt      1      1     11
  1982-Reagan.txt      1      1     17
  1983-Reagan.txt      3      7     19
  1984-Reagan.txt      3      5     23
  1985-Reagan.txt      1      1     12
  1986-Reagan.txt      2      2     14
  1987-Reagan.txt      1      0     24
  1988-Reagan.txt      1      0     16
    1989-Bush.txt      2      3     13
    1990-Bush.txt      3      2      9
  1991-Bush-1.txt      2      2     13
  1991-Bush-2.txt      7      7     13
    1992-Bush.txt      4      4     26
 1993-Clinton.txt      1      2     45
 1994-Clinton.txt      1      1     63
 1995-Clinton.txt      1      3     73
 1996-Clinton.txt      2      3     40
 1997-Clinton.txt      1      2     30
 1998-Clinton.txt      2      2     22
 1999-Clinton.txt      2      3     22
 2000-Clinton.txt      5      6     41
2001-GWBush-1.txt      3      3     14
2001-GWBush-2.txt      1      2     12
  2002-GWBush.txt      3      5     14
  2003-GWBush.txt      6      4     33
  2004-GWBush.txt      6      8     21
  2005-GWBush.txt      8     11     18
  2006-GWBush.txt      7      7     22
```

b) To determine what happens with the usage of the words over time, I first collected all the possible years that are documented, in a list. Then I created a conditional frequency distribution, using the words as experiments, and the years (the first 4 characters of the filename) as the condition. For each file that begins with the year I want to check out, I iterate over the words for each word I am looking for, in the same way as I did with the counts. Then I used the plot() function which shows that the number of the word “men” has been pretty stable, except for the year 1965. The word “people” peaked in 1995 and 1946, with a lot of swinging otherwise. There have also been more occurrences of “women” over the time, especially since 1991.



Exercise 3

a) The converter to Pig Latin is pretty straightforward and can be seen clearly in the code itself, so I am not going to explain it here. However, I did use an additional function syllables() which I have borrowed from Stack Overflow (link in the source code). The function finds the syllable count of a word, which helps me convert longer words that begin with vowels. My code fails in one case: if a word begins with a vowel sound, like the word “honest”. I am not sure how to fix this, maybe it is possible to look up the phonetic version of the word and go from there.

b) For converting text to Pig Latin, I used my function from a) to convert each word. First, however, I find the indices of all the characters in the text, to be able to reconstruct it correctly with all the punctuation marks and whitespace characters that would get deleted by the tokenizer I use to filter out the actual words. Basically, I convert each word in the text to Pig Latin and concatenate it to the already converted words. After concatenating, I check whether the next index contains a punctuation mark and/or whitespace characters by checking the char_indices list, and grabbing those from the original text, if needed. However, my code does not work for apostrophes because the tokenizer splits on apostrophes. Also, the word “am” will not convert for some reason, because the syllable count function does not show the correct number of syllables in “am”.

c) Explanation from source code:
“First, we have to remove all the "ay" and "yay" from the words. For words than begin with vowels, we're done after having removed "yay" (unless the alternative encoding is used, where the first vowel + consonant cluster is removed). Then we have to distinguish between words beginning with one consonant or a consonant cluster. The only way I can think of is to try both cases (i.e., moving the last letter to the first place and moving the two last letters to the first place) and checking if any of the words make sense, which would be very time consuming, at least if we are using Wikipedia's definitions of Pig Latin.

Ex. keyboard -> erboardkay and training -> ainingtray would become either keyboard/dkeyboar and rainingt/training.

The computer would not know any better which one is correct. We would also have to account for words that start with three consonants, which would add even more ambiguity to the task. I do not know how to do this without human intervention or some kind of thesaurus that is available for the computer”

Exercise 4

I used the code provided to us on GitLab to set up Selenium and get all the “tabs” containing Reddit posts. For each tab I extracted the subreddit, upvotes, title, and time posted:

```
r/dankmemes Tue Feb 2 08:12:55 2021 UTC 45.0k Rule 34 goes brrrrrr

r/wallstreetbets Tue Feb 2 10:03:49 2021 UTC 17.8k It's not a loss if you don't sell, Poland holding
strong! Just eat some pierogi and relax. 💎🙌

r/wallstreetbets Tue Feb 2 05:28:56 2021 UTC 33.2k Mark Cuban AMA at 9:30 AM CST, Tuesday 2/1/2021

r/Music Mon Feb 1 21:59:19 2021 UTC 16.6k Marilyn Manson Dropped by Record Label After Abuse
Allegations

r/wallstreetbets Tue Feb 2 11:26:37 2021 UTC 9328 To all GME holders: Shut up and listen

r/nba Tue Feb 2 02:37:13 2021 UTC 12.6k [Highlight] LeBron gets heckled by Karen

r/wallstreetbets Tue Feb 2 07:25:51 2021 UTC 38.5k So You're Experiencing FUD

r/gaming Mon Feb 1 18:50:07 2021 UTC 160k Happy Black History Month

r/LeopardsAteMyFace Mon Feb 1 17:04:00 2021 UTC 50.9k Horn-man wants revenge on Trump for not
pardoning him after following his marching orders

r/Cringetopia Mon Feb 1 21:22:07 2021 UTC 46.0k Am i the only one who thinks YOKO ONO is the queen of
cringe?

r/wallstreetbets Tue Feb 2 11:00:23 2021 UTC 1994 Daily Discussion Thread for February 02, 2021

r/MurderedByAOC Mon Feb 1 19:21:22 2021 UTC 28.8k AOC torches Biden: "$2,000 means $2,000. $2,000
does not mean $1,400."
```

Exercise 5

a) I used the boilerplate code provided to us on GitLab to set up Tweepy. I chose the following topics: games, food, cats, nature, education, computers, love, Trump, Biden, and Norway. For each topic, I gathered tweets using a query, while filtering out retweets to reduce the amount of work to be done later. I gathered English tweets only, saving them inside a .txt file for each category, of the format tweets_[category].txt. I used a Categorized Plaintext Reader to read the corpus by category induced from the file names. I also set up a list of stopwords, which is partly made up of stopwords from this website https://sites.google.com/site/iamgongwei/home/sw and partly my own chosen stopwords. I chose to use stopword list made specifically for Twitter because it is domain specific and therefore more effective than a general stopword list, such as the one from NLTK. I used a Plaintext Corpus Reader to read the corpus of stopwords within the corpus of Tweets.
I tokenized the Tweets by taking the raw text of the whole category file and running it through various functions that removed links, emojis (borrowed from GitHub, link in source code), tagged users and numbers. I made two versions of a tokenized Tweet, one with hashtags and one without. I used NLTK's Tweet Tokenizer for both versions, in combination with the stopword list. The version without hashtags will be used to determine the most common words in the corpus.

b) I made a frequency distribution of all the tokenized tweets without hashtags to find the 10 most common words in the corpus:

```
[('nature', 95), ('education', 87), ('computers', 84), ('biden', 71), ('trump', 66),
('norway', 66), ('games', 51), ('baekhyun', 39), ('exo', 39), ('unity', 33)]
```

c) I also made a frequency distribution of the tokenized tweets from each category to find the 10 most common words from each category, pasted here for convenience. I am getting ". \r\n" as a word in multiple places, and I have tried to fix it with no avail. I hope it does not ruin my report too much:

```
Most common words in biden : [('biden', 56), ('trump', 19), ('governing', 7), ('joe', 6),
("biden's", 5), ('president', 5), ('abortion', 4), ('administration', 4), ('energy', 4),
('fact', 4)]

Most common words in cats : [('cats', 26), ('cat', 7), ('lisa', 5), ('catsofinstagram', 4),
('. \r\n.', 4), ('catlover', 3), ('catstagram', 3), ('bbb', 3), ('catlovers', 3), ('sao', 3)]

Most common words in computers : [('computers', 84), ('computer', 9), ('.\r\n.', 6),
('laptops', 6), ('data', 6), ('personal', 6), ('software', 6), ('laptop', 5), ('library', 5),
('technology', 5)]

Most common words in education : [('education', 71), ('nature', 16), ('science', 16),
('covidexplained', 15), ('unity', 15), ('wisdom', 15), ('15', חינוך), ('spirituality', 14),
('covid', 10), ('children', 6)]

Most common words in food : [('eating', 4), ('junk', 3), ('army', 3), ('talking', 2),
('croissant', 2), ('bakery', 2), ('thinking', 2), ('server', 2), ('logs', 2), ('weight', 2)]

Most common words in games : [('games', 50), ('game', 8), ('utd', 7), ('playing', 4),
('podcast', 4), ('drummond', 4), ('allen', 4), ('ortg', 4), ('drtg', 4), ('hill', 3)]
```

```
Most common words in love : [('baekhyun', 39), ('exo', 38), ('breath', 5), ('piece', 5),
('song', 3), ('umk', 2), ('attention', 2), ('weloveyouharryfromiran', 2), ('kairo', 2),
('shinee', 2)]
```

```
Most common words in nature : [('nature', 78), ('education', 15), ('science', 14),
('covidexplained', 14), ('unity', 14), ('wisdom', 14), ('חינוך', 14), ('spirituality', 13),
('covid', 10), ('jihoon', 9)]
```

```
Most common words in norway : [('norway', 66), ('uk', 14), ('eu', 8), ('wealth', 7), ('fund',
7), ('canada', 7), ('mining', 6), ('sweden', 6), ('spain', 6), ('sovereign', 4)]
```

```
Most common words in trump : [('trump', 46), ('biden', 13), ('president', 9), ('election', 6),
('gop', 4), ('fact', 4), ('lost', 4), ('russia', 4), ('republicans', 4), ("trump's", 4)]
```

d) I did the same thing here, only this time I used the hash tagged version of the corpus, and filtered out the hashtags only:

```
Most common hashtags in education : [('#education', 24), ('#covidexplained', 15), ('#nature',
15), ('#unity', 15), ('#wisdom', 15), ('#חינוך', 15), ('#science', 14), ('#spirituality', 14),
('#covid', 10), ('#budget', 3)]
```

```
Most common hashtags in nature : [('#nature', 25), ('#covidexplained', 14), ('#education',
14), ('#unity', 14), ('#wisdom', 14), ('#חינוך', 14), ('#science', 13), ('#spirituality', 13),
('#covid', 10), ('#treasure', 3)]
```

```
Most common hashtags in computers : [('#computers', 7), ('#trustintech', 2), ('#technology',
2), ('#laptops', 2), ('#computer', 2), ('#laptop', 2), ('#linux', 2), ('#training', 2),
('#ubuntu', 2), ('#windows', 2)]
```

```
Most common hashtags in trump : [('#biden', 2), ('#israël', 2), ('#ישראל', 2), ('#pp', 1),
('#dv', 1), ('#onev', 1), ('#ovelections', 1), ('#traitortuberville', 1), ('#gopcomplicit',
1), ('#gopcorruption', 1)]
```

```
Most common hashtags in norway : [('#norway', 13), ('#covid', 2), ('#france', 2), ('#spain',
2), ('#germany', 2), ('#tomyonlyxiu', 2), ('#ferrynews', 1), ('#watercolor', 1),
('#watercolour', 1), ('#akvarell', 1)]
```

```
Most common hashtags in love : [('#baekhyun', 38), ('#exo', 38), ('#umk', 2), ('#___', 2),
('#sidharthhitsmontwitter', 2), ('#shinee', 2), ('#weloveyouharryfromiran', 2), ('#hbd_xydo',
1), ('#candy', 1), ('#slove', 1)]
```

```
Most common hashtags in games : [('#games', 4), ('#gaming', 2), ('#videogames', 2), ('#gamer',
2), ('#podcast', 2), ('#videopodcast', 2), ('#podcasting', 2), ('#talk', 2), ('#talkshow', 2),
('#ps', 2)]
```

```
Most common hashtags in food : [('#food', 4), ('#unitedkingdom', 2), ('#singapore', 2),
('#mewgulfmusicfesbyud', 2), ('#ateez', 1), ('#farmers', 1), ('#ਲੜਾਂਗੇ', 1), ('#no', 1),
('#croissantday', 1), ('#happycroissantday', 1)]
```

```
Most common hashtags in cats : [('#cats', 8), ('#cat', 4), ('#catsofinstagram', 4),
('#catlover', 3), ('#catstagram', 3), ('#catlovers', 3), ('#animals', 2), ('#pets', 2),
('#catsoftwitter', 2), ('#catlife', 2)]
```

```
Most common hashtags in biden : [('#biden', 3), ('#demvoice', 2), ('#trump', 2),
('#muslimban', 2), ('#votethemout', 1), ('#jigyasaquiz', 1), ('#tangentiaquiz', 1),
('#bluepath', 1), ('#wtpblue', 1), ('#wtpeduon', 1)]
```