# Week 3 - Text Analysis & Working with Text Files

(Be sure to copy to drive)

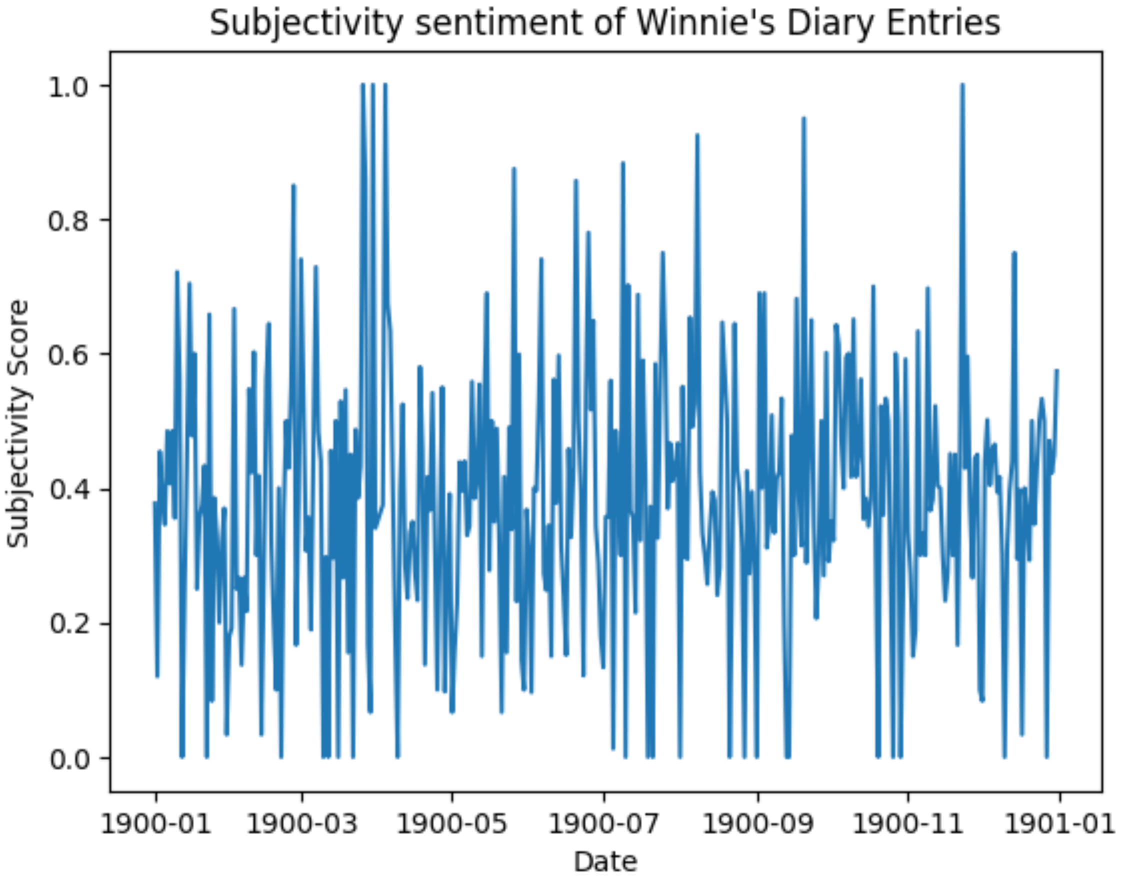
Text data is a bit different from numeric data. We can easily find the average of a series of numbers and things like the highest and lowest values in a range to get some ideas on what we are dealing with. We can't really do that with text. We'll focus on some tools that you can use to actually analyze text. We'll start with a library called TextBlob.

```
#Load up our libraries
from textblob import TextBlob
from google.colab import drive

#these should look familar
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import requests

#Some extra libraries we'll need for text analysis
import nltk
nltk.download('punkt')
nltk.download('brown')
nltk.download('punkt_tab')


#Connect to Gdrive
drive.mount('/content/gdrive')

print("Libraries and Drive Ready!")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Unzipping corpora/brown.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
Mounted at /content/gdrive
Libraries and Drive Ready!
```

**Q2** Modify the following code cell to create a line graph of the *subjectivity* of her diary entries for the year.

```
plt.plot(winnie_corpus["date"],winnie_corpus["subjectivity"])
plt.title("Subjectivity sentiment of Winnie's Diary Entries")
plt.ylabel("Subjectivity Score")
plt.xlabel("Date")
plt.show()
```



## Noun Phrases for the Diary

Now let's generate the noun phrases for January's entries

```
#we use some pandas work to just grab the January entries
#stretch back into your memory to think about conditionals again
jan_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-01-01') & (winnie_

jan_phrases = dict()

for entry in jan_corpus.entry:

    tb = TextBlob(entry)
    #we create a dictionary that will hold the noun phrases
    #if it is the first time we see this np we put it in the dictionary
    #if not, we must have a count already, so we increase that by one
    for np in tb.noun_phrases:
        if np in jan_phrases:
            jan_phrases[np] += 1
            jan_phrases[np] += 1
        else:
            jan_phrases[np] = 1

#Print the top 10 things she mentioned in January

for np in sorted(jan_phrases, key=jan_phrases.get, reverse=True)[0:10]:
    print(np, jan_phrases[np])


```

```
went 23
mamma 19
papa 16
beatrice 8
took 8
trusty 5
music lesson 4
mr read 4
annie gardiner 4
got 4
```

**Q5** Modify the next series of cells to generate noun phrases for the next 5 months of the year.

```
#February Entries
feb_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-02-01') & (winnie_

feb_phrases = dict()

for entry in feb_corpus.entry:
    tb = TextBlob(entry)
    for np in tb.noun_phrases:
        if np in feb_phrases:
            feb_phrases[np] += 1
        else:
            feb_phrases[np] = 1

#Print the top 10 things she mentioned in February

for np in sorted(feb_phrases, key=feb_phrases.get, reverse=True)[0:10]:
    print(np, feb_phrases[np])
```

```
mamma 20
went 18
papa 12
beatrice 7
pay 6
was 5
took 5
got 5
music lesson 4
```

```
musical 4
```

```python
#March Entries
mar_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-03-01') & (winnie_


mar_phrases = dict()

for entry in mar_corpus.entry:
    tb = TextBlob(entry)
    for np in tb.noun_phrases:
        if np in mar_phrases:
            mar_phrases[np] += 1
        else:
            mar_phrases[np] = 1

#Print the top 10 things she mentioned in March

for np in sorted(mar_phrases, key=mar_phrases.get, reverse=True)[0:10]:
    print(np, mar_phrases[np])
```

```
mamma 21
took 9
papa 9
went 8
got 7
pay 7
dwyer 7
mr perry 6
godard 5
beatrice 5
```

```python
#April Entries
april_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-04-01') & (winni

april_phrases = dict()

for entry in april_corpus.entry:
    tb = TextBlob(entry)
    for np in tb.noun_phrases:
        if np in april_phrases:
            april_phrases[np] += 1
        else:
            april_phrases[np] = 1

#Print the top 10 things she mentioned in April

for np in sorted(april_phrases, key=april_phrases.get, reverse=True)[0:10]:
    print(np, april_phrases[np])
```

```
mamma 24
mamma 24
papa 20
went 11
got 9
pay 8
took 7
beatrice 7
sullivan 7
christ 6
lizzie 5
```

```
#May Entries
may_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-05-01') & (winnie_

may_phrases = dict()

for entry in may_corpus.entry:
    tb = TextBlob(entry)
    for np in tb.noun_phrases:
        if np in may_phrases:
            may_phrases[np] += 1
        else:
            may_phrases[np] = 1

#Print the top 10 things she mentioned in may

for np in sorted(may_phrases, key=may_phrases.get, reverse=True)[0:10]:
    print(np, may_phrases[np])
```

```
mamma 30
papa 21
went 17
got 11
ella 7
lizzie 6
rode 5
helen 5
pay 4
dwyer 4
```

```
#June Entries
june_corpus = winnie_corpus[(winnie_corpus['date'] >= '1900-06-01') & (winnie_

june_phrases = dict()

for entry in june_corpus.entry:
    tb = TextBlob(entry)
    for np in tb.noun_phrases:
        if np in june_phrases:
            june_phrases[np] += 1
        else:
            june_phrases[np] = 1
            june_phrases[np] = 1

#Print the top 10 things she mentioned in june

for np in sorted(june_phrases, key=june_phrases.get, reverse=True)[0:10]:
    print(np, june_phrases[np])
```

```
mamma 21
papa 16
got 13
went 13
took 7
lizzie 7
was 4
mrs. sullivan 4
ella 4
aunt lillie 4
```