Exercise 7 Text Analysis

Skills

- learn to write regular expressions (re)

Resources

- https://docs.python.org/3/library/re.html
- https://www.dataquest.io/blog/regex-cheatsheet/
- http://www.pyregex.com/

In [1]:

```
# import modules
import re, os
import pandas as pd
import numpy as np
```

Part A: Regular Expressions

Pattern Matching

Here are some strings to practice creating regular expressions to match.

In [5]:

```
# write a regular expression to match dates of the form MM/DD/YYYY
pattern01 = ['01/21/2018', '12/12/2012', '03/03/2018']
for x in pattern01:
    print(re.search('\d{2}/\d{2}/\d{2}', x))
# loop through pattern01
# use your regular expression to search for a match
# print the output to verify it matched the example pattern
# example:
# for x in pattern_object:
#    print(INSERT REGEX SYNTAX HERE)
```

```
<re.Match object; span=(0, 8), match='01/21/20'>
<re.Match object; span=(0, 8), match='12/12/20'>
<re.Match object; span=(0, 8), match='03/03/20'>
```

In [22]:

```
# write a regular expression to match dates of the form Month Day, Year
pattern02 = ['March 8, 2017', 'January 15, 2018', 'May 3, 2017']
for x in pattern02:
    print(re.search('\w+ \d+, \d{4}', x))
# loop through pattern02
# use your regular expression to search for a match
# print the output to verify it matched the example pattern
```

```
<re.Match object; span=(0, 13), match='March 8, 2017'>
<re.Match object; span=(0, 16), match='January 15, 2018'>
<re.Match object; span=(0, 11), match='May 3, 2017'>
```

In [16]:

```
#  write a regular expression to match Email addresses of the form username@host
pattern03 = ['email@umd.edu', 'email@terpmail.umd.edu', 'some.email@ox.ac.uk']
for x in pattern03:
    print(re.search('\w+\@\w+.\w+.\w+', x))
# loop through pattern03
# use your regular expression to search for a match
# print the output to verify it matched the example pattern
```

```
<re.Match object; span=(0, 13), match='email@umd.edu'>
<re.Match object; span=(0, 22), match='email@terpmail.umd.edu'>
<re.Match object; span=(5, 19), match='email@ox.ac.uk'>
```

In [6]:

```
#  write a regular expression to match Social Security Numbers
pattern04 = ['111-11-1111', '999-19-1919', '888-12-3434']
for x in pattern04:
    print(re.search('\d{3}-\d{2}-\d{4}', x))
# loop through pattern04
# use your regular expression to search for a match
# print the output to verify it matched the example pattern
```

```
<re.Match object; span=(0, 11), match='111-11-1111'>
<re.Match object; span=(0, 11), match='999-19-1919'>
<re.Match object; span=(0, 11), match='888-12-3434'>
```

In [ ]: