Data-X Spring 2019: Homework 7

Webscraping

In this homework, you will do some exercises with web-scraping.

Name: Shrey Samdani

SID: 303200414

Fun with Webscraping & Text manipulation

1. Statistics in Presidential Debates

Your first task is to scrape Presidential Debates from the Commission of Presidential Debates website: https://www.debates.org/voter-education/debate-transcripts/ (https://www.debates.org/voter-education/debate-transcripts/)

To do this, you are not allowed to manually look up the URLs that you need, instead you have to scrape them. The root url to be scraped is the one listed above, namely: https://www.debates.org/voter-education/debate-transcripts/ (https://www.debates.org/voter-education/debate-transcripts/)

1. By using `requests` and `BeautifulSoup` find all the links / URLs on the website that links to transcriptions of **First Presidential Debates** from the years [1988, 1984, 1976, 1960]. In total you should find 4 links / URLs that fulfill this criteria. **Print the urls.**
2. When you have a list of the URLs your task is to create a Data Frame with some statistics (see example of output below):
   A. Scrape the title of each link and use that as the column name in your Data Frame.
   B. Count how long the transcript of the debate is (as in the number of characters in transcription string). Feel free to include `\` characters in your count, but remove any breakline characters, i.e. `\n` . You will get credit if your count is +/- 10% from our result.
   C. Count how many times the word **war** was used in the different debates. Note that you have to convert the text in a smart way (to not count the word **warranty** for example, but counting **war.**, **war!**, **war,** or **War** etc.
   D. Also scrape the most common used word in the debate, and write how many times it was used. Note that you have to use the same strategy as in C in order to do this.

   Print your final output result.

Tips:

In order to solve the questions above, it can be useful to work with Regular Expressions and explore methods on strings like `.strip()`, `.replace()`, `.find()`, `.count()`, `.lower()` etc. Both are very powerful tools to do string processing in Python. To count common words for example I used a `Counter` object and a Regular expression pattern for only words, see example:

```
from collections import Counter
import re

counts = Counter(re.findall(r"[\w']+", text.lower()))
```

Read more about Regular Expressions here: https://docs.python.org/3/howto/regex.html (https://docs.python.org/3/howto/regex.html)

Example output of all of the answers to Question 1.2:

	September 25, 1988: The First Bush-Dukakis Presidential Debate	[redacted]	[redacted]	[redacted]
Debate char length	87488	[redacted]	[redacted]	[redacted]
war_count				
most_common_w	[redacted]	[redacted]	[redacted]	[redacted]
most_common_w_count	[redacted]	[redacted]	[redacted]	[redacted]

.

In [1]:
```
import requests
import bs4 as bs

dates = [1988, 1984, 1976, 1960]
source = requests.get("https://www.debates.org/voter-education/debate-trans
soup = bs.BeautifulSoup(source.content, features='html.parser')
links = []
titles = []
for link in soup.find_all("a"):
    strLink = str(link)
    for date in dates:
        if str(date) in strLink and "First" in strLink:
            titles.append(link.contents[0])
            links.append("https://www.debates.org"+strLink.split('"')[1])
            print(links[-1])
```

```
https://www.debates.org/voter-education/debate-transcripts/september-25-1
988-debate-transcript/ (https://www.debates.org/voter-education/debate-tr
anscripts/september-25-1988-debate-transcript/)
https://www.debates.org/voter-education/debate-transcripts/october-7-1984
-debate-transcript/ (https://www.debates.org/voter-education/debate-trans
cripts/october-7-1984-debate-transcript/)
https://www.debates.org/voter-education/debate-transcripts/september-23-1
976-debate-transcript/ (https://www.debates.org/voter-education/debate-tr
anscripts/september-23-1976-debate-transcript/)
https://www.debates.org/voter-education/debate-transcripts/september-26-1
960-debate-transcript/ (https://www.debates.org/voter-education/debate-tr
anscripts/september-26-1960-debate-transcript/)
```

In [2]:

```
import pandas as pd
import re
from collections import Counter
from collections import defaultdict

data = defaultdict(list)

for i in range(4):
    source = requests.get(links[i])
    content = source.text.split(r"</strong>")[2]
    soup = bs.BeautifulSoup(content, features='html.parser')

    length = 0
    text = ""
    for p in soup.find_all("p"):
        length += len(p.text)
        text +=" " + p.text
    data[titles[i]].append(length)

    words = Counter(re.findall(r"[\w']+", text.lower()))
    data[titles[i]].append(words['war'])

    data[titles[i]].append(max(words, key=words.get))
    data[titles[i]].append(words[data[titles[i]][-1]])


df = pd.DataFrame(data, index=["Debate char length","war_count","most_commo
df
```

Out[2]:

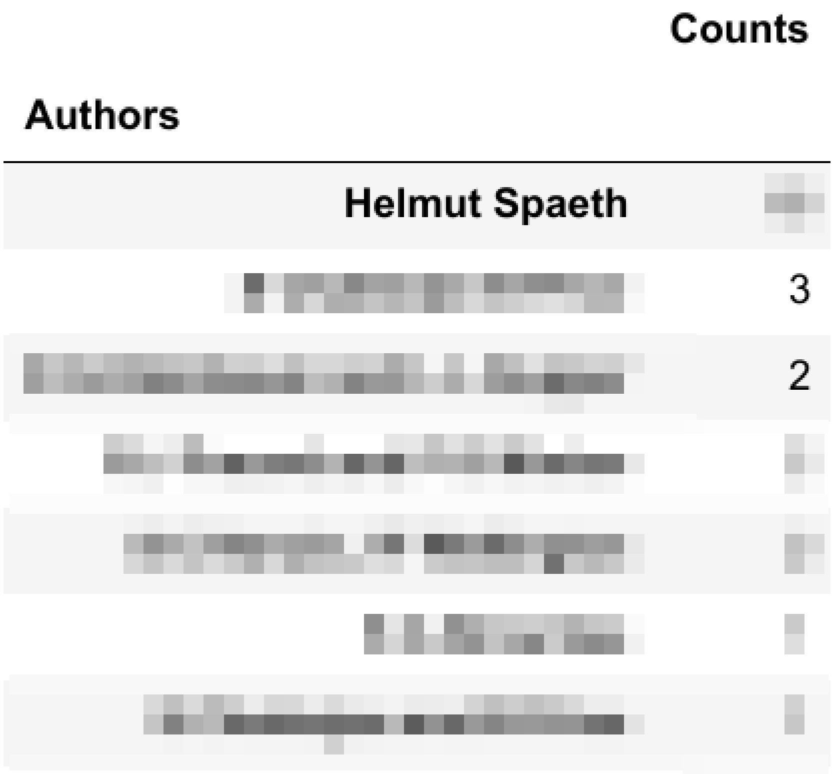
	September 25, 1988: The First Bush-Dukakis Presidential Debate	October 7, 1984: The First Reagan-Mondale Presidential Debate	September 23, 1976: The First Carter-Ford Presidential Debate	September 26, 1960: The First Kennedy-Nixon Presidential Debate
Debate char length	87469	86490	80717	60917
war_count	8	2	7	3
most_common_w	the	the	the	the
most_common_w_count	804	867	857	779

2. Download and read in specific line from many data sets

Scrape the first 27 data sets from this URL http://people.sc.fsu.edu/~jburkardt/datasets/regression/ (http://people.sc.fsu.edu/~jburkardt/datasets/regression/) (i.e. `x01.txt - x27.txt`). Then, save the 5th line in each data set, this should be the name of the data set author (get rid of the `#` symbol, the white spaces and the comma at the end).

Count how many times (with a Python function) each author is the reference for one of the 27 data sets. Showcase your results, sorted, with the most common author name first and how many times he appeared in data sets. Use a Pandas DataFrame to show your results, see example. **Print your final output result.**

Example output of the answer for Question 2:



	Counts
Authors	
Helmut Spaeth	[redacted]
[redacted]	3
[redacted]	2
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

```
In [3]: url = "http://people.sc.fsu.edu/~jburkardt/datasets/regression/"
        counts = defaultdict(int)
        for i in range(1,28):
            source = requests.get(url+"x%02d.txt" % i).text
            counts[source.split("\n")[4][5:-1]] += 1
```

```
In [4]: df = pd.DataFrame(list(counts.items()), columns = ["Authors","Counts"])
        df.sort_values(by="Counts", ascending = False)
```

```
Out[4]:
                      Authors  Counts
0               Helmut Spaeth      16
5        S Chatterjee, B Price       3
1     R J Freund and P D Minton     2
2  D G Kleinbaum and L L Kupper     2
6     S C Narula, J F Wellington     2
3                 K A Brownlee      1
4       S Chatterjee and B Price     1
```