Scraping Wikipedia

Question 1: Individual page scraping

There are in total 5676 featured articles, 21.79% (1237) of them are biographies.

I determine which articles are biographies by first finding out all the h3 tags. Then I find all titles under h3 tags that have ‘mw-headline’ and ‘biographies’.

Question 2: Scraping a dataset

99.76% of first paragraphs were scraped.
The first paragraph of each biography has been written in file first_paragraphs.txt

I failed to scrape some pages because I couldn’t retrieve them.

Question 3: Extracting information from messy content

My approach is that, for each biography, I count the frequencies of pronouns that appear in the biography and retrieve the maximum of them. For example, if male, female, plural pronouns appear respectively 10, 5, 2 times, then male is believed to be the pronoun of the biography. The drawback of this approach is that when the frequencies of pronouns are very close to each other, this approach may choose the wrong pronoun. Furthermore, the pronoun of the character of the biography may not even appear the most frequently in the biography because the page may use a lot of words to describe another person/thing related to the character. All in all, this approach may fail often because it does not consider the context of each pronoun. Also, through this approach, I excluded content that does not have a pronoun, uses “it” as the pronoun, or describes a number of persons all together.

The percentage of biographies of each pronoun are as follows.

84.88% biographies use he/his pronouns.
12.29% biographies use she/her pronouns.
1.94% biographies use they/them pronouns.
0.00% biographies use unknown pronouns.
Failed to parse 0.89% of pages.

I failed to parse 0.89% of pages because I couldn’t retrieve them.
Based on my approach, no biography uses unknown pronouns because all biographies use a gender pronoun somewhere in the page for at least one time.

Question 4: Additional Analysis

What is the average number of words used to describe the object of a biography?
For a biography to be “described in detail”, how many words should be used to describe it?
For a biography to be “described in brief”, how many words should be used to describe it?

These questions are interesting because the extent of details that a biography covers, which is measured by the number of words, could be an important metric for evaluating a biography.

The distribution of the word counts are as follows.

max: 36018, min: 80, mean: 8748.12, median: 7403.5, std: 5326.18

The result shows that it is appropriate to use 7000 – 9000 words for a biography, based on the mean and median. If a biography exceeds 14000 words (mean + std), then this biography is counted as “described in detail”. If a biography has fewer than 3500 words (mean – std), then this biography is counted as “described in brief”.

Question 5: Preparing a dataset for sharing

The data frame is saved to export_dataset.csv

There are three columns in the csv file. They are:
title: the biography page title
pronoun: the most common pronoun of the biography.
len: the number of words in the biography.

I failed to scrape 11 pages because I couldn’t retrieve them. The limitation on future analysis is that the future study will not be able to study all biography pages and see the full picture.

Other scientists could use my code to retrieve the titles of all the biography pages or all the non-biography pages, if they are interested. They could also use my code to retrieve the first paragraphs of all pages and adapt the code to retrieve specific paragraph of interest. They could also use my code to find the most common pronouns of each page and adapt my code to retrieve other characteristics, such as the most common profession.

When using my code, the data scientists should make sure that they have the wiki_api file. They should also import the following packages and function before running the code.

```
import re
import pandas as pd
from wiki_api import page_text
from bs4 import BeautifulSoup
import numpy as np
```

Sample code that allows future users to load the file:

```
    pd.read_csv(export_dataset.csv)
```

Basic analysis on length to confirm that the dataset is successfully downloaded:

```
    file_len = 0
    with open(“export_dataset.csv”) as export_dataset:
    for row in export_dataset:
        file_len += 1
```

```
if file_len == 1226:
        print (“the dataset is successfully downloaded”)
else:
        print(“the dataset is not successfully downloaded”)
```

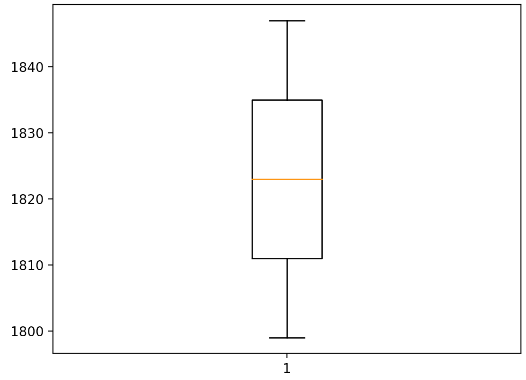
Part 3: Extra Credit

5.17% of featured biographies describe people who are currently alive.

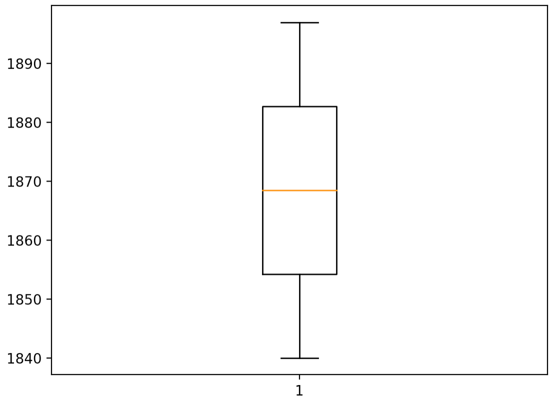
78.09% featured biographies are unknown. I failed to find data on these biographies because I failed to scrape some pages and some pages do not have the birth and death date.

The distribution of dates of birth and death on featured biographies is in the following boxplots. It shows that the featured biographies sites mainly include people born in the early 1980s. People who already passed away mainly died in the late 1980s.

The result is saved to extra_credit.csv



Boxplot of Birth Date



Boxplot of Death Date