Info

Name: Jacqui Unciano **Date:** 1/22/24 **Assignment:** HW01

Directions

Using the notebook we reviewed in class (M01_03_first_foray.ipynb) as your guide, create a notebook to import the text contained the attached file (pg42324.txt Download pg42324.txt) as a data frame of lines (not chunks). Once you have done this, answer these questions or perform the task listed. In your notebook, create a section for each question.vels.

In [2]:
```
import pandas as pd
import configparser
```

In [3]:
```
config = configparser.ConfigParser()
config.read("../../../env.ini")
data_home = config['DEFAULT']['data_home']
output_dir = config['DEFAULT']['output_dir']
```

In [4]:
```
data_home, output_dir
```

Out[4]:
```
('/Users/jacqu/OneDrive/Documents/MSDS-at-UVA-2023/DS5001/data',
 '/Users/jacqu/OneDrive/Documents/MSDS-at-UVA-2023/DS5001/output')
```

In [11]:
```
src_file = f"{data_home}/pg42324.txt"
lines = open(src_file, 'r').readlines()
text = pd.DataFrame(lines)
text.columns = ['line_str']
text.index.name = 'line_num'
```

In [12]:
```
text.head()
```

Out[12]:

line_num	line_str
0	ï»¿The Project Gutenberg EBook of Frankenstein...
1	\n
2	This eBook is for the use of anyone anywhere a...
3	almost no restrictions whatsoever. You may co...
4	re-use it under the terms of the Project Guten...

Question 1

File failed to load: /extensions/MathZoom.js

How many tokens does the raw text have? By raw text, we mean the text as-is, without all of the Gutenberg boilerplate removed.

In [13]:
```
token_df = text.line_str.str.split(expand=True).stack().to_frame('token_str')
token_df.index.names = ['line_num','token_num']
token_df.head()
```

Out[13]:

		token_str
line_num	**token_num**	
0	**0**	ï»¿The
	1	Project
	2	Gutenberg
	3	EBook
	4	of

In [14]:
```
token_df.shape
```

Out[14]:
```
(80985, 1)
```

The number of tokens the raw text has is 80,985.

Question 2

What is the most frequent pronoun in the text?

In [15]:
```
token_df['term_str'] = token_df.token_str.str.replace(r'\W+', '', regex=True).str.l
token_df.head()
```

Out[15]:

		token_str	term_str
line_num	**token_num**		
0	**0**	ï»¿The	ïthe
	1	Project	project
	2	Gutenberg	gutenberg
	3	EBook	ebook
	4	of	of

In [16]:
```
vocab_df = token_df.term_str.value_counts().to_frame('n')
vocab_df.index.name = 'term_str'
```

In [18]:
```
vocab_df.head(10)
```

File failed to load: /extensions/MathZoom.js

Out[18]:

term_str	n
the	4574
and	3120
i	2918
of	2918
to	2257
my	1819
a	1497
in	1232
was	1064
that	1060

In this text, the most frequent pronoun is "i".

Question 3&4

Which subject pronoun is most frequent in the text we imported in class? Provide a brief explanation for this difference, based on what you may know about the two novels.

I believe the pronoun most frequent in the Sense and Sensibility text was 'her'. Sense and Sensibility is a novel featuring the lives of two sisters whereas Frankenstein is a novel that details about a monster learning about himself. Since the Jane Austin novel is about two women, it only makes sense that the most frequent pronoun is 'she' (that, plus from what I can remember, the novel is written in 3rd person POV). That, compared to Frankenstein, which is more about what Frankenstein's monster is thinking about and experiencing, the difference makes sense (that, and I'm pretty sure is written in 1st person POV--I could go into great detail about the implications and objective that Mary Shelly had when putting the novelin 1st person, but I won't).

File failed to load: /extensions/MathZoom.js