Module 1 HW

Kurt Samuels Jr

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.

In [1]:
```
import pandas as pd
import configparser
```

In [3]:
```
config = configparser.ConfigParser()
config.read("/Users/kurtsamuels/Desktop/msds/spring/DS5001/env.ini")
data_home = config['DEFAULT']['data_home']
output_dir = config['DEFAULT']['output_dir']
```

In [4]:
```
src_file = f"{data_home}/pg42324.txt"
```

In [7]:
```
lines = open(src_file, 'r').readlines()

lines[:5]
```

Out[7]:
```
['\ufeffThe Project Gutenberg EBook of Frankenstein, by Mary W. Shelley\n',
 '\n',
 'This eBook is for the use of anyone anywhere at no cost and with\n',
 'almost no restrictions whatsoever.  You may copy it, give it away or\n',
 're-use it under the terms of the Project Gutenberg License included\n']
```

In [8]:
```
text = pd.DataFrame(lines)
text
```

Out[8]:

	0
0	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...
...	...
8023	\n
8024	This Web site includes information about Proje...
8025	including how to make donations to the Project...
8026	Archive Foundation, how to help produce our ne...
8027	subscribe to our email newsletter to hear abou...

8028 rows × 1 columns

1.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 [25]:
```
with open(src_file, 'r', encoding='utf-8') as file:
    raw_text = file.read()
tokens = raw_text.split()
total_tokens = len(tokens)
print(f"The raw text contains {total_tokens} tokens.")
```

```
The raw text contains 80985 tokens.
```

2. What is the most frequent pronoun in the text?

In [27]:
```
pronouns = ['i', 'you', 'he', 'she', 'it', 'we', 'they', 'me', 'him', 'her',
normalized_tokens = [token.lower() for token in tokens]
pronoun_counts = {pronoun: normalized_tokens.count(pronoun) for pronoun in p

# Create a DataFrame from the counts
pronoun_df = pd.DataFrame(list(pronoun_counts.items()), columns=['Pronoun',
pronoun_df
```

Out[27]:

	Pronoun	Count
0	i	2794
1	you	512
2	he	596
3	she	242
4	it	466
5	we	184
6	they	212
7	me	483
8	him	131
9	her	321
10	us	36
11	them	75
12	my	1798
13	your	254
14	his	550
15	hers	2
16	its	133
17	our	137
18	their	190

I is the pronoun that appears the most

3. Which subject pronoun is most frequent in the text we imported in class?

In [17]:
```
subject_pronouns = ['i', 'he', 'she', 'we', 'they','you', 'it']
```

In [28]:
```
src_file2 = f"{data_home}/gutenberg/pg105.txt"
lines2 = open(src_file2, 'r').readlines()
with open(src_file2, 'r', encoding='utf-8') as file:
    raw_text2 = file.read()
tokens2 = raw_text2.split()
normalized_tokens2 = [token.lower() for token in tokens2]
subpronoun_counts = {subpronoun: normalized_tokens2.count(subpronoun) for su
# Create a DataFrame from the counts
subpronoun_df = pd.DataFrame(list(subpronoun_counts.items()), columns=['Subj
subpronoun_df
```

Out[28]:

	Subject Pronoun	Count
0	i	986
1	he	917
2	she	1113
3	we	152
4	they	427
5	you	555
6	it	825

4. Provide a brief explanation for this difference, based on what you may know about the two novels.

The prevelance of the personal pronoun I in Frankenstein suggests that it is told from a first person perspective, while Persuasion is told from a third person perspective as seen from its heavy use of third person subject pronouns such as she, he and they. She being the most prevalent as the story follows a female lead.