# Ian Yung

## *1.21.2024*

## *m01hw.ipynb*

# Set Up

In [35]:

```
import pandas as pd
```

In [36]:

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

# Import File

In [37]:

```
src_file = f"{data_home}/HW1/pg42324.txt"
```

In [38]:

```
lines = open(src_file, 'r').readlines()
```

In [39]:

```
lines[:5]
```

Out[39]:

```
['\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']
```

# Convert to Dataframe

In [40]:

```
text = pd.DataFrame(lines)
```

In [41]:

```
text
```

Out[41]:

|  | 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

## Question 1

In [42]:
```
chunk_pat = '\n\n'
```

In [43]:
```
chunks = open(src_file, 'r').read().split(chunk_pat)
```

In [44]:
```
text = pd.DataFrame(chunks, columns=['chunk_str'])
text.index.name = 'chunk_id'
```

In [45]:
```
text.chunk_str = text.chunk_str.str.replace('\n+', ' ', regex=True).str.stri
```

In [46]:
```
text.head()
```

Out[46]:

|  | chunk_str |
|---|---|
| chunk_id |  |
| 0 | The Project Gutenberg EBook of Frankenstein, ... |
| 1 | This eBook is for the use of anyone anywhere a... |
| 2 | Title: Frankenstein or, The Modern Prom... |
| 3 | Author: Mary W. Shelley |
| 4 | Release Date: March 13, 2013 [EBook #42324] |

## Convert lines to tokens

In [47]:

```
K = text.chunk_str.str.split(expand=True).stack().to_frame('token_str')
K.index.names = ['chunk_num','token_num']
```

In [48]:

```
K
```

Out[48]:

| | | token_str |
|---|---|---|
| **chunk_num** | **token_num** | |
| **0** | **0** | The |
| | **1** | Project |
| | **2** | Gutenberg |
| | **3** | EBook |
| | **4** | of |
| **...** | **...** | ... |
| **941** | **35** | to |
| | **36** | hear |
| | **37** | about |
| | **38** | new |
| | **39** | eBooks. |

80985 rows × 1 columns

There are 80,985 tokens of the raw text.

# Question 2

In [49]:

```
K['term_str'] = K.token_str.str.replace(r'\W+', '', regex=True).str.lower()
```

In [50]:

```
V = K.term_str.value_counts().to_frame('n')
V.index.name = 'term_str'
```

In [51]:

```
V.head(10)
```

Out[51]:

| term_str | n |
|---|---|
| the | 4575 |
| and | 3120 |
| of | 2918 |
| i | 2918 |
| to | 2257 |
| my | 1819 |
| a | 1497 |
| in | 1232 |
| was | 1064 |
| that | 1060 |

"I" is the most prominent pronoun.

# Question 3

In [52]:
```
src_file = f"{data_home}/gutenberg/pg105.txt"
chunks = open(src_file, 'r').read().split(chunk_pat)
```

In [53]:
```
text = pd.DataFrame(chunks, columns=['chunk_str'])
text.index.name = 'chunk_id'
text.chunk_str = text.chunk_str.str.replace('\n+', ' ', regex=True).str.stri
```

In [54]:
```
K = text.chunk_str.str.split(expand=True).stack().to_frame('token_str')
K.index.names = ['chunk_num','token_num']
```

In [55]:
```
K['term_str'] = K.token_str.str.replace(r'\W+', '', regex=True).str.lower()
```

In [56]:
```
V = K.term_str.value_counts().to_frame('n')
V.index.name = 'term_str'
```

In [57]:
```
V.head(10)
```

Out[57]:

|  | n |
|---|---|
| term_str |  |
| the | 3501 |
| to | 2862 |
| and | 2851 |
| of | 2684 |
| a | 1648 |
| in | 1439 |
| was | 1336 |
| her | 1202 |
| had | 1187 |
| she | 1143 |

"She" is the most prominent subject pronoun of the Jane Austen novel.

## Question 4

Based on my level of knowledge of the two authors and their work, Frankenstein is a story centrally focused on a "mad scientist" and his creation, while Jane Austen's work tends to revolve around the lives of young aristocratic women in 19th century British society. It seems fairly intuitive that, within this context, "she" will feature prominently within *Persuasion* by Jane Austen, since the novel follows the lives of several women, while *Frankenstein*, which is narrated in first-person, is obviously going to rely upon "I" often throughout the story.