HW: First Foray

Course: DS 5001 Exploratory Text Analytics
Module: 01 Getting Started
Topic: HW: First Foray
Author: Lindley Slipetz
Date: 18 January 2025

In [3]:
```
import pandas as pd
```

In [4]:
```
import configparser
config = configparser.ConfigParser()
```

In [7]:
```
config.read("../../../env.ini.ini")
```

Out[7]:
```
[]
```

In [8]:
```
src_file = "C:\\Users\\ddj6tu\\Documents\\GitHub\\DS5001\\data\\pg42324.txt"
```

In [9]:
```
lines = open(src_file, 'r').readlines()
```

In [10]:
```
text = pd.DataFrame(lines)
text
```

Out[10]:

	0
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...
...	...
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

In [11]:
```
chunk_pat = '\n\n'
```

In [12]:
```
chunks = open(src_file, 'r').read().split(chunk_pat)
```

In [13]:
```
text = pd.DataFrame(chunks, columns=['chunk_str'])
text.index.name = 'chunk_id'
```

In [14]:
```
text.head()
```

Out[14]:

chunk_id	chunk_str
0	ï»¿The Project Gutenberg EBook of Frankenstein...
1	This eBook is for the use of anyone anywhere a...
2	\nTitle: Frankenstein\n or, The Modern P...
3	Author: Mary W. Shelley
4	Release Date: March 13, 2013 [EBook #42324]

In [15]:
```
text.chunk_str = text.chunk_str.str.replace('\n+', ' ', regex=True).str.strip()
```

In [16]:
```
K = text.chunk_str.str.split(expand=True).stack().to_frame('token_str')
K.index.names = ['chunk_num','token_num']
```

In [18]:
```
K
```

Out[18]:

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

80985 rows × 1 columns

Question 1: There are 80,985 tokens.

In [20]:
```
K['term_str'] = K.token_str.str.replace(r'\W+', '', regex=True).str.lower()
V = K.term_str.value_counts().to_frame('n')
V.index.name = 'term_str'
```

In [21]:
```
V.head(10)
```

Out[21]:

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

Question 2: "I" is the most frequent pronoun.

Question 3: "She" was the most frequent pronoun in the example we did in class.

Question 4: It's been a long time since I've read these books, but I would guess that Frankenstein is written in the first person and Persuasion is written in the third person.