M01 Homework

Michael Vaden, mtv2eva

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.

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

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

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

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

In [ ]:
```
!ls -l {data_home}
```

```
total 1880
-rw-r--r--  1 michaelvaden  staff  494987 Jan 21 20:39 pg105.txt
-rw-r--r--  1 michaelvaden  staff  465627 Jan 21 13:32 pg42324.txt
```

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

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

t = pd.DataFrame(lines, columns=['line_str'])

K = t.line_str.str.split(expand=True).stack().to_frame('token_str')
K.index.names = ['line_num','token_num']

K['token_str'] = K['token_str'].str.lower()

K
```

Out[ ]:

		token_str
line_num	token_num	
0	0	the
	1	project
	2	gutenberg
	3	ebook
	4	of
...	...	...
8027	5	to
	6	hear
	7	about
	8	new
	9	ebooks.

80985 rows × 1 columns

There are **80985** tokens in the raw text

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

In [ ]:
```
V = K.token_str.value_counts().to_frame('n')

V.head()
```

Out[ ]:

	n
the	4539
and	3002
of	2913
i	2794
to	2244

Without even needing to do a specific search for pronouns, we can see that **I** is the most frequent pronoun in the text

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

In [ ]:
```
src_file2 = f"{data_home}/pg105.txt"
```

In [ ]:
```
chunk_pat = '\n\n'
chunks2 = open(src_file2, 'r').read().split(chunk_pat)

text2 = pd.DataFrame(chunks2, columns=['chunk_str'])
text2.index.name = 'chunk_id'

text2.chunk_str = text2.chunk_str.str.replace('\n+', ' ', regex=True).str.strip()

K2 = text2.chunk_str.str.split(expand=True).stack().to_frame('term_str')
K2.index.names = ['chunk_num','token_num']

K2['term_str'] = K2.term_str.str.replace(r'\W+', '', regex=True).str.lower()

V2 = K2.term_str.value_counts().to_frame('n')

V2.query("index in ['I', 'you', 'he', 'she', 'it', 'we', 'they']")
```

Out[ ]:

	n
she	1143
it	1051
he	963
you	695
they	435
we	163

As we can see above, **she** is the most frequent subject pronoun

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

In *Persuasion*, by Jane Austen, the author often writes in 3rd person. This provides an explanation for why **she** is the most common pronoun in the novel, where there are many female characters. However, in *Frankenstein*, by Mary Shelley, the author writes more in a first person narrative, which is a reasonable explanation as to why **I** is the most common pronoun in the novel.