Assignment 2.16 Code Exercise 2

Max Ryoo (hr2ee)

Q1

Download this Gutenberg version of Jane Austen's Sense and Sensibility.

The data is in the subfolder data with the name pg161.txt

In [1]:

```
import pandas as pd
%matplotlib inline
sense_sensibility = "data/pg161.txt"
```

Q2

Write a Jupyter Notebook to convert the raw text into a data frame of tokens, just as we did with Persuasion.

- You may use the Code Walkthrough notebook, which is attached to this assignment, as your template.

Be sure to include the OHCO of Chapters, Paragraphs, and Sentences in your dataframe's index.

First read in the data as a dataframe.

In [2]:

```
epub = open(sense_sensibility, 'r', encoding='utf-8-sig').readlines()
df = pd.DataFrame(epub, columns=['line_str'])
df.index.name = 'line_num'
df.line_str = df.line_str.str.strip()
```

In [3]:

```
df.head(10)
```

Out[3]:

line_num	line_str
0	The Project Gutenberg EBook of Sense and Sensi...
1	
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...
5	with this eBook or online at www.gutenberg.net
6	
7	
8	Title: Sense and Sensibility
9	

Extract the title from the first line

In [4]:
```
title = df.loc[0].line_str.replace('The Project Gutenberg EBook of ', '')
df['title'] = title
```

In [5]:
```
print(title)
```

```
Sense and Sensibility, by Jane Austen
```

In [6]:
```
df.head()
```

Out[6]:

line_num	line_str	title
0	The Project Gutenberg EBook of Sense and Sensi...	Sense and Sensibility, by Jane Austen
1		Sense and Sensibility, by Jane Austen
2	This eBook is for the use of anyone anywhere a...	Sense and Sensibility, by Jane Austen
3	almost no restrictions whatsoever. You may co...	Sense and Sensibility, by Jane Austen
4	re-use it under the terms of the Project Guten...	Sense and Sensibility, by Jane Austen

Clean up the data by removing the Gutenberg's Front and back matter

In [7]:
```
start = df.line_str.str.match(r"\*\*\*\s*START OF (THE|THIS) PROJECT")
end = df.line_str.str.match(r"\*\*\*\s*END OF (THE|THIS) PROJECT")

start_index = df.loc[start].index[0]
end_index = df.loc[end].index[0]
```

In [8]:
```
"The novel starts on line {} and ends on line {}".format(start_index, end_index
```

Out[8]:
```
'The novel starts on line 18 and ends on line 12668'
```

In [9]:

```
df.loc[start_index:end_index]
```

Out[9]:

	line_str	title
line_num		
18	*** START OF THIS PROJECT GUTENBERG EBOOK SENS...	Sense and Sensibility, by Jane Austen
19		Sense and Sensibility, by Jane Austen
20	Special thanks are due to Sharon Partridge for...	Sense and Sensibility, by Jane Austen
21	proofreading and correction of this etext.	Sense and Sensibility, by Jane Austen
22		Sense and Sensibility, by Jane Austen
...	...	...
12664		Sense and Sensibility, by Jane Austen
12665		Sense and Sensibility, by Jane Austen
12666	End of the Project Gutenberg EBook of Sense an...	Sense and Sensibility, by Jane Austen
12667		Sense and Sensibility, by Jane Austen
12668	*** END OF THIS PROJECT GUTENBERG EBOOK SENSE ...	Sense and Sensibility, by Jane Austen

12651 rows × 2 columns

We see from this indexed dataframe that techicanlly the novel doesn't start at the given indexs. We will need to subset the beginning for the thanks of the book by the author and the ending whitespaces.

In [10]:

```
df.loc[start_index:end_index].head(20)
```

Out[10]:

	line_str	title
line_num		
18	*** START OF THIS PROJECT GUTENBERG EBOOK SENS...	Sense and Sensibility, by Jane Austen
19		Sense and Sensibility, by Jane Austen
20	Special thanks are due to Sharon Partridge for...	Sense and Sensibility, by Jane Austen
21	proofreading and correction of this etext.	Sense and Sensibility, by Jane Austen
22		Sense and Sensibility, by Jane Austen
23		Sense and Sensibility, by Jane Austen
24		Sense and Sensibility, by Jane Austen
25		Sense and Sensibility, by Jane Austen
26		Sense and Sensibility, by Jane Austen
27		Sense and Sensibility, by Jane Austen
28		Sense and Sensibility, by Jane Austen
29		Sense and Sensibility, by Jane Austen
30		Sense and Sensibility, by Jane Austen
31		Sense and Sensibility, by Jane Austen
32		Sense and Sensibility, by Jane Austen
33	SENSE AND SENSIBILITY	Sense and Sensibility, by Jane Austen
34		Sense and Sensibility, by Jane Austen
35	by Jane Austen	Sense and Sensibility, by Jane Austen
36		Sense and Sensibility, by Jane Austen
37	(1811)	Sense and Sensibility, by Jane Austen

In [11]:

```
df.loc[start_index:end_index].tail(20)
```

Out[11]:

	line_str	title
line_num		
12649	merits and the happiness of Elinor and Mariann...	Sense and Sensibility, by Jane Austen
12650	as the least considerable, that though sisters...	Sense and Sensibility, by Jane Austen
12651	within sight of each other, they could live wi...	Sense and Sensibility, by Jane Austen
12652	between themselves, or producing coolness betw...	Sense and Sensibility, by Jane Austen
12653		Sense and Sensibility, by Jane Austen
12654		Sense and Sensibility, by Jane Austen
12655		Sense and Sensibility, by Jane Austen
12656	THE END	Sense and Sensibility, by Jane Austen
12657		Sense and Sensibility, by Jane Austen
12658		Sense and Sensibility, by Jane Austen
12659		Sense and Sensibility, by Jane Austen
12660		Sense and Sensibility, by Jane Austen
12661		Sense and Sensibility, by Jane Austen
12662		Sense and Sensibility, by Jane Austen
12663		Sense and Sensibility, by Jane Austen
12664		Sense and Sensibility, by Jane Austen
12665		Sense and Sensibility, by Jane Austen
12666	End of the Project Gutenberg EBook of Sense an...	Sense and Sensibility, by Jane Austen
12667		Sense and Sensibility, by Jane Austen
12668	*** END OF THIS PROJECT GUTENBERG EBOOK SENSE ...	Sense and Sensibility, by Jane Austen

In [12]:

```
df = df.loc[start_index+15:end_index-12]
df
```

Out[12]:

	line_str	title
line_num		
33	SENSE AND SENSIBILITY	Sense and Sensibility, by Jane Austen
34		Sense and Sensibility, by Jane Austen
35	by Jane Austen	Sense and Sensibility, by Jane Austen
36		Sense and Sensibility, by Jane Austen
37	(1811)	Sense and Sensibility, by Jane Austen
...	...	...
12652	between themselves, or producing coolness betw...	Sense and Sensibility, by Jane Austen
12653		Sense and Sensibility, by Jane Austen
12654		Sense and Sensibility, by Jane Austen
12655		Sense and Sensibility, by Jane Austen
12656	THE END	Sense and Sensibility, by Jane Austen

12624 rows × 2 columns

Now that we have the actual contents of the novel, we will need to include the OHCO of Chapters, Paragraphs, Sentences.

In [13]:

```
OHCO = ['chap_num', 'para_num', 'sent_num', 'token_num']
```

First lets fill in the Chapter information from chucking the chapters.

In [14]:

```
chap_lines = df.line_str.str.match(r"^\s*(chapter|letter)\s+(\d+)", case=False)
```

In [15]:

```
df.loc[chap_lines]
```

Out[15]:

	line_str	title
line_num		
42	CHAPTER 1	Sense and Sensibility, by Jane Austen
196	CHAPTER 2	Sense and Sensibility, by Jane Austen
399	CHAPTER 3	Sense and Sensibility, by Jane Austen
562	CHAPTER 4	Sense and Sensibility, by Jane Austen
757	CHAPTER 5	Sense and Sensibility, by Jane Austen
859	CHAPTER 6	Sense and Sensibility, by Jane Austen
987	CHAPTER 7	Sense and Sensibility, by Jane Austen
1113	CHAPTER 8	Sense and Sensibility, by Jane Austen
1245	CHAPTER 9	Sense and Sensibility, by Jane Austen
1449	CHAPTER 10	Sense and Sensibility, by Jane Austen
1666	CHAPTER 11	Sense and Sensibility, by Jane Austen
1817	CHAPTER 12	Sense and Sensibility, by Jane Austen
1998	CHAPTER 13	Sense and Sensibility, by Jane Austen
2282	CHAPTER 14	Sense and Sensibility, by Jane Austen
2441	CHAPTER 15	Sense and Sensibility, by Jane Austen
2719	CHAPTER 16	Sense and Sensibility, by Jane Austen
2946	CHAPTER 17	Sense and Sensibility, by Jane Austen
3154	CHAPTER 18	Sense and Sensibility, by Jane Austen
3332	CHAPTER 19	Sense and Sensibility, by Jane Austen
3633	CHAPTER 20	Sense and Sensibility, by Jane Austen
3914	CHAPTER 21	Sense and Sensibility, by Jane Austen
4215	CHAPTER 22	Sense and Sensibility, by Jane Austen
4533	CHAPTER 23	Sense and Sensibility, by Jane Austen
4768	CHAPTER 24	Sense and Sensibility, by Jane Austen
5002	CHAPTER 25	Sense and Sensibility, by Jane Austen
5198	CHAPTER 26	Sense and Sensibility, by Jane Austen
5455	CHAPTER 27	Sense and Sensibility, by Jane Austen
5733	CHAPTER 28	Sense and Sensibility, by Jane Austen
5884	CHAPTER 29	Sense and Sensibility, by Jane Austen
6325	CHAPTER 30	Sense and Sensibility, by Jane Austen
6629	CHAPTER 31	Sense and Sensibility, by Jane Austen
7005	CHAPTER 32	Sense and Sensibility, by Jane Austen
7279	CHAPTER 33	Sense and Sensibility, by Jane Austen
7602	CHAPTER 34	Sense and Sensibility, by Jane Austen

	line_str	title
line_num		
7889	CHAPTER 35	Sense and Sensibility, by Jane Austen
8153	CHAPTER 36	Sense and Sensibility, by Jane Austen
8457	CHAPTER 37	Sense and Sensibility, by Jane Austen
8901	CHAPTER 38	Sense and Sensibility, by Jane Austen
9206	CHAPTER 39	Sense and Sensibility, by Jane Austen
9409	CHAPTER 40	Sense and Sensibility, by Jane Austen
9707	CHAPTER 41	Sense and Sensibility, by Jane Austen
9978	CHAPTER 42	Sense and Sensibility, by Jane Austen
10156	CHAPTER 43	Sense and Sensibility, by Jane Austen
10491	CHAPTER 44	Sense and Sensibility, by Jane Austen
11061	CHAPTER 45	Sense and Sensibility, by Jane Austen
11279	CHAPTER 46	Sense and Sensibility, by Jane Austen
11572	CHAPTER 47	Sense and Sensibility, by Jane Austen
11839	CHAPTER 48	Sense and Sensibility, by Jane Austen
11987	CHAPTER 49	Sense and Sensibility, by Jane Austen
12411	CHAPTER 50	Sense and Sensibility, by Jane Austen

Assign numbers to chapters

```
In [16]: chap_nums = [i+1 for i in range(df.loc[chap_lines].shape[0])]
```

```
In [17]: df.loc[chap_lines, 'chap_num'] = chap_nums
```

```
In [18]: df.chap_num = df.chap_num.ffill()
```

```
In [19]: df
```

Out[19]:

	line_str	title	chap_num
line_num			
33	SENSE AND SENSIBILITY	Sense and Sensibility, by Jane Austen	NaN
34		Sense and Sensibility, by Jane Austen	NaN
35	by Jane Austen	Sense and Sensibility, by Jane Austen	NaN
36		Sense and Sensibility, by Jane Austen	NaN
37	(1811)	Sense and Sensibility, by Jane Austen	NaN
...	...	...	...
12652	between themselves, or producing coolness betw...	Sense and Sensibility, by Jane Austen	50.0
12653		Sense and Sensibility, by Jane Austen	50.0
12654		Sense and Sensibility, by Jane Austen	50.0
12655		Sense and Sensibility, by Jane Austen	50.0
12656	THE END	Sense and Sensibility, by Jane Austen	50.0

12624 rows × 3 columns

However, we see that the title and author and heading lines have NaN as the chap_num. We will need to clean these up by removing these heading lines.

In [20]:

```
df = df.loc[~df.chap_num.isna()] # Remove chapter heading lines
df = df.loc[~chap_lines] # Remove everything before Chapter 1
df.chap_num = df.chap_num.astype('int') # Convert chap_num from float to int
```

In [21]:

```
df.sample(10)
```

Out[21]:

line_num	line_str	title	chap_num
11360	resolute firmness, as if determined at once to...	Sense and Sensibility, by Jane Austen	46
3067	"Perhaps, then, you would bestow it as a rewar...	Sense and Sensibility, by Jane Austen	17
5076	every inconvenience of that kind, should disre...	Sense and Sensibility, by Jane Austen	25
612	But of his minuter propensities, as you call t...	Sense and Sensibility, by Jane Austen	4
8444	life, as she was with them; had given each of ...	Sense and Sensibility, by Jane Austen	36
12582	valued friend; and to see Marianne settled at ...	Sense and Sensibility, by Jane Austen	50
2776		Sense and Sensibility, by Jane Austen	16
7667	had their assiduities made them to her, that t...	Sense and Sensibility, by Jane Austen	34
8140		Sense and Sensibility, by Jane Austen	35
3777	Parliament!--won't it? How I shall laugh! It ...	Sense and Sensibility, by Jane Austen	20

Grouping the lines by chapter num

In [22]:
```
dfc = df.groupby(OHCO[:1]).line_str.apply(lambda x: '\n'.join(x)).to_frame() #
```

In [23]:
```
dfc.head()
```

Out[23]:

chap_num	line_str
1	\n\nThe family of Dashwood had long been settl...
2	\n\nMrs. John Dashwood now installed herself m...
3	\n\nMrs. Dashwood remained at Norland several ...
4	\n\n"What a pity it is, Elinor," said Marianne...
5	\n\nNo sooner was her answer dispatched, than ...

Now that we have chapters, we should move onto paragraphs.

In [24]:
```
dfp = dfc['line_str'].str.split(r'\n\n+', expand=True).stack()\
    .to_frame().rename(columns={0:'para_str'})
```

In [25]:
```
dfp.head()
```

Out[25]:

chap_num		para_str
1	0	
	1	The family of Dashwood had long been settled i...
	2	By a former marriage, Mr. Henry Dashwood had o...
	3	The old gentleman died: his will was read, and...
	4	Mr. Dashwood's disappointment was, at first, s...

Renaming the second index as para_num (paragraph number)

In [26]:
```
dfp.index.names = OHCO[:2]
dfp.head()
```

Out[26]:

chap_num	para_num	para_str
1	0	
	1	The family of Dashwood had long been settled i...
	2	By a former marriage, Mr. Henry Dashwood had o...
	3	The old gentleman died: his will was read, and...
	4	Mr. Dashwood's disappointment was, at first, s...

We will still need to do more cleaning by removing empty paragraphs

In [27]:
```
dfp['para_str'] = dfp['para_str'].str.replace(r'\n', ' ').str.strip()
dfp = dfp[~dfp['para_str'].str.match(r'^\s*$')] # Remove empty paragraphs
```

```
/var/folders/pn/dgy7ckd90nl7mlj6g6rc_1kw0000gn/T/ipykernel_42550/3422525513.p
y:1: FutureWarning: The default value of regex will change from True to False
in a future version.
  dfp['para_str'] = dfp['para_str'].str.replace(r'\n', ' ').str.strip()
```

In [28]:
```
dfp.head()
```

Out[28]:

chap_num	para_num	para_str
1	1	The family of Dashwood had long been settled i...
	2	By a former marriage, Mr. Henry Dashwood had o...
	3	The old gentleman died: his will was read, and...
	4	Mr. Dashwood's disappointment was, at first, s...
	5	His son was sent for as soon as his danger was...

Now that we have chapter and paragraph information, we will split into sentences.

```
In [29]: dfs = dfp['para_str'].str.split(r'[.?!;:"]+', expand=True).stack()\
             .to_frame().rename(columns={0:'sent_str'})
```

```
In [30]: dfs.index.names = OHCO[:3]
         dfs = dfs[~dfs['sent_str'].str.match(r'^\s*$')] # Remove empty paragraphs
```

```
In [31]: dfs.head()
```

Out[31]:

			sent_str
chap_num	**para_num**	**sent_num**	
1	**1**	**0**	The family of Dashwood had long been settled i...
		1	Their estate was large, and their residence ...
		2	The late owner of this estate was a single m...
		3	But her death, which happened ten years befo...
		4	for to supply her loss, he invited and receiv...

The final splitting will be into tokens.

```
In [32]: dft = dfs['sent_str'].str.split(r"[\s',-]+", expand=True).stack()\
             .to_frame().rename(columns={0:'token_str'})
```

```
In [33]: dft.index.names = OHCO[:4]
```

```
In [34]: dft.head()
```

Out[34]:

				token_str
chap_num	**para_num**	**sent_num**	**token_num**	
1	**1**	**0**	**0**	The
			1	family
			2	of
			3	Dashwood
			4	had

Combining the title information will make us have one dataframe with a complete break down of the novel.

```
In [35]: df_sense = dft
         df_sense['title'] = title
```

```
In [36]: df_sense = df_sense.reset_index()
         df_sense
```

Out[36]:

	chap_num	para_num	sent_num	token_num	token_str	title
0	1	1	0	0	The	Sense and Sensibility, by Jane Austen
1	1	1	0	1	family	Sense and Sensibility, by Jane Austen
2	1	1	0	2	of	Sense and Sensibility, by Jane Austen
3	1	1	0	3	Dashwood	Sense and Sensibility, by Jane Austen
4	1	1	0	4	had	Sense and Sensibility, by Jane Austen
...	...	...	...	...	...	...
127744	50	21	1	42	between	Sense and Sensibility, by Jane Austen
127745	50	21	1	43	their	Sense and Sensibility, by Jane Austen
127746	50	21	1	44	husbands	Sense and Sensibility, by Jane Austen
127747	50	22	0	0	THE	Sense and Sensibility, by Jane Austen
127748	50	22	0	1	END	Sense and Sensibility, by Jane Austen

127749 rows × 6 columns

Q3

Once you have done this, try to extend the notebook to combine both Persuasion and Sense and Sensibility into a single dataframe with an appropriately modified OHCO list.

- In other words, make sure your index includes a level (column) for the book title. You may want to export the dataframe created in the first notebook to a CSV file and then import it into the second notebook as a dataframe.
- See attached files for Persuasion.

From the sample notebook, our final output was given by the austen-persuassion.csv file.

In [37]:

```
df_persuasion = pd.read_csv('austen-persuasion.csv')
df_persuasion.head()
```

Out[37]:

	chap_num	para_num	sent_num	token_num	token_str
0	1	1	0	0	Sir
1	1	1	0	1	Walter
2	1	1	0	2	Elliot
3	1	1	0	3	of
4	1	1	0	4	Kellynch

We will need to add a title column for the same column names as our dataframe for sense and sensibility

In [38]:

```
df_persuasion['title'] = 'Persuasion, by Jane Austen'
df_persuasion
```

Out[38]:

	chap_num	para_num	sent_num	token_num	token_str	title
0	1	1	0	0	Sir	Persuasion, by Jane Austen
1	1	1	0	1	Walter	Persuasion, by Jane Austen
2	1	1	0	2	Elliot	Persuasion, by Jane Austen
3	1	1	0	3	of	Persuasion, by Jane Austen
4	1	1	0	4	Kellynch	Persuasion, by Jane Austen
...	...	...	...	...	...	...
88944	24	14	0	6	of	Persuasion, by Jane Austen
88945	24	14	0	7	Persuasion	Persuasion, by Jane Austen
88946	24	14	0	8	by	Persuasion, by Jane Austen
88947	24	14	0	9	Jane	Persuasion, by Jane Austen
88948	24	14	0	10	Austen	Persuasion, by Jane Austen

88949 rows × 6 columns

In [39]:

```
merged = pd.concat([df_sense, df_persuasion])
```

In [40]:

```
merged
```

Out[40]:

	chap_num	para_num	sent_num	token_num	token_str	title
0	1	1	0	0	The	Sense and Sensibility, by Jane Austen
1	1	1	0	1	family	Sense and Sensibility, by Jane Austen
2	1	1	0	2	of	Sense and Sensibility, by Jane Austen
3	1	1	0	3	Dashwood	Sense and Sensibility, by Jane Austen
4	1	1	0	4	had	Sense and Sensibility, by Jane Austen
...	...	...	...	...	...	...
88944	24	14	0	6	of	Persuasion, by Jane Austen
88945	24	14	0	7	Persuasion	Persuasion, by Jane Austen
88946	24	14	0	8	by	Persuasion, by Jane Austen
88947	24	14	0	9	Jane	Persuasion, by Jane Austen
88948	24	14	0	10	Austen	Persuasion, by Jane Austen

216698 rows × 6 columns

In [ ]: