Assignment 1.15

Max Ryoo (hr2ee)

Set up

In [5]:
```
import pandas as pd
%matplotlib inline
```

In [4]:
```
epub_file = "2701-0.txt"
```

Import into DataFrame

In [6]:
```
epub = open(epub_file, '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 [7]:
```
df.sample(10)
```

Out[7]:

line_num	line_str
18059	(_Carpenter standing before his vice-bench, an...
6724	topmaul: “a white whale. Skin your eyes for hi...
1567	little more in my shirt sleeves. But beginning...
1291	to our lips cups of scalding tea with our half...
19271	is that stove? In the stern-sheets, man; where...
14651	between. In no living thing are the lines of b...
7168	dancing girls!—the Heeva-Heeva! Ah! low veiled...
4427	wonderingly looking from me to Queequeg, with ...
12672	whale and the ship—where he would occasionally...
12641	any way get rid of the dangerous liabilities w...

Questions

1. How many lines does the file have?
2. At what line number does the novel’s contents begin?
3. At what line number does the novel’s content end?

Q1

How many lines does the file have?

In [9]:
```
df.shape
```

Out[9]:
```
(22316, 1)
```

From the text to pandas dataframe procedure, we can see that the number of lines is 22316. When reading in the dataframe, a new row was created for each new line, which is why we can state the the shape of the dataframe will define how many lines there were in the text file.

Q2

At what line number does the novel's contents begin?

In [26]:
```
start_boiler = df.line_str.str.match(r"\*\*\*\s*START OF (THE|THIS) PROJECT")
start_line = df.loc[start_boiler].index[0]
start_line
```

Out[26]:
```
23
```

In [28]:
```
df.loc[start_line : start_line + 10]
```

Out[28]:

line_num	line_str
23	*** START OF THE PROJECT GUTENBERG EBOOK MOBY-...
24	
25	
26	
27	
28	MOBY-DICK;
29	
30	or, THE WHALE.
31	
32	By Herman Melville
33	

The start of this project marks that start of the content. Therefore, the actual start as shown in the above dataframe can be seen as line number 24 (after the start comment)

Q3

At what line number does the novel's content end?

In [29]:
```
end_boiler = df.line_str.str.match(r"\*\*\*\s*END OF (THE|THIS) PROJECT")
```

```
end_line = df.loc[end_boiler].index[0]
end_line
```

Out[29]:
```
21964
```

In [30]:
```
df.loc[end_line - 10 : end_line]
```

Out[30]:

line_num	line_str
21954	one whole day and night, I floated on a soft a...
21955	unharming sharks, they glided by as if with pa...
21956	the savage sea-hawks sailed with sheathed beak...
21957	sail drew near, nearer, and picked me up at la...
21958	devious-cruising Rachel, that in her retracing...
21959	children, only found another orphan.
21960	
21961	
21962	
21963	
21964	*** END OF THE PROJECT GUTENBERG EBOOK MOBY-DI...

Similar to the above start finding procedure, the ending comment is found on line number 21964. Theoretically the content of the book will end on 21964 - 1 = 21963.

The full contents is shown below

In [32]:
```
df.loc[start_line+1: end_line-1]
```

Out[32]:

line_num	line_str
24	
25	
26	
27	
28	MOBY-DICK;
...	...
21959	children, only found another orphan.
21960	
21961	
21962	
21963	

21940 rows × 1 columns

In [ ]: