Project 6: Sentiment Scoring

CS 5473: Data Mining Fall 2022

Instructor: Dr. Mohammad Imran Chowdhury

Total Points: 75

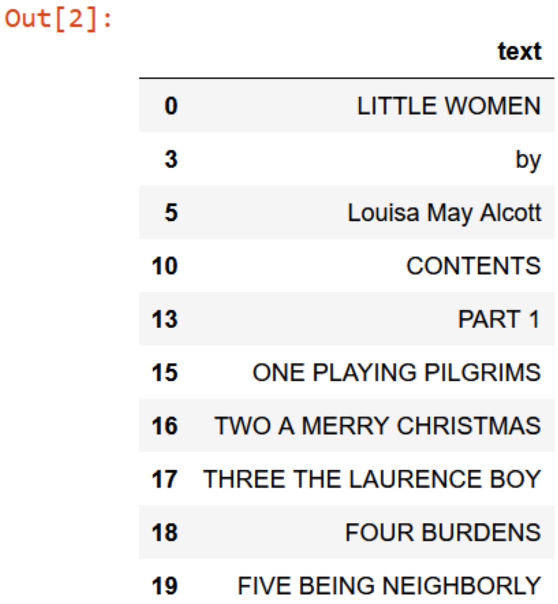
Due: 12/01/2022 11:59 PM

In this project, I invite you to do the following:

1. Import and prepare the text `LittleWomen.txt` dataset
2. Tokenize the data.
3. Score the sentiments.
4. Calculate average sentiment scores for each section of 100 lines.
5. Graph the "sentiment arc" of the story.

Task 1: Import the text `LittleWomen.txt` dataset **(10 points)**

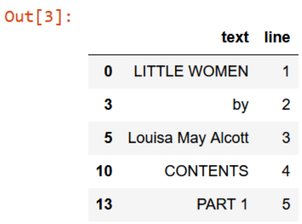
Load the dataset `LittleWomen.txt` provided to you as **'data/LittleWomen.txt'** file into the Jupyter Notebook. Show the first 10 rows. The output should be as follows: **(5 points)**



Out[2]:

	text
0	LITTLE WOMEN
3	by
5	Louisa May Alcott
10	CONTENTS
13	PART 1
15	ONE PLAYING PILGRIMS
16	TWO A MERRY CHRISTMAS
17	THREE THE LAURENCE BOY
18	FOUR BURDENS
19	FIVE BEING NEIGHBORLY

Next, Add Line Numbers. The output should be as follows for the first 5 rows: **(5 points)**

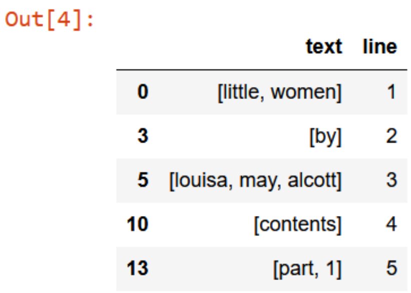


Out[3]:

	text	line
0	LITTLE WOMEN	1
3	by	2
5	Louisa May Alcott	3
10	CONTENTS	4
13	PART 1	5

Task 2: Tokenize the data. **(15 points)**

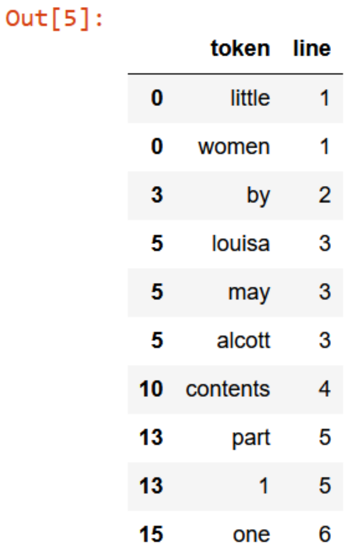
Tokenize the Data. The output should be as follows if you show first 5 rows: **(10 points)**



Out[4]:

	text	line
0	[little, women]	1
3	[by]	2
5	[louisa, may, alcott]	3
10	[contents]	4
13	[part, 1]	5

Now, Collect Tokens into a Single Series. The output should be as follows if you show first 10 rows: **(5 points)**

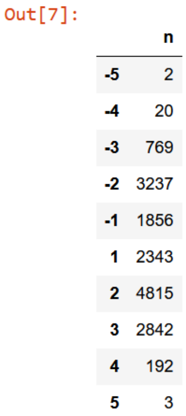


Out[5]:

	token	line
0	little	1
0	women	1
3	by	2
5	louisa	3
5	may	3
5	alcott	3
10	contents	4
13	part	5
13	1	5
15	one	6

Task 3: Score the sentiments. **(20 points)**

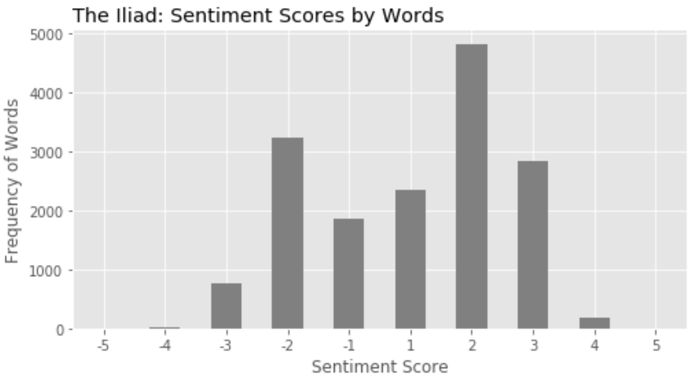
Calculate sentiment scores using the AFINN lexicon, which scores words on a scale of -5 (most negative) to +5 (most positive). And, show a frequency table for the sentiment scores. **(10 points)**



Out[7]:

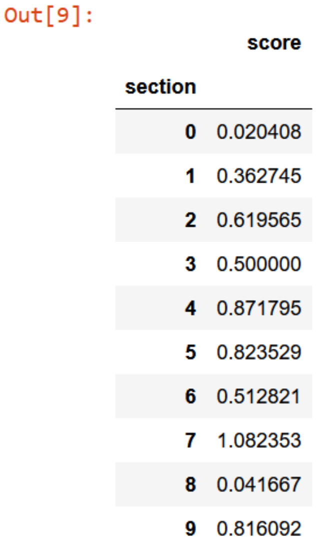
	n
-5	2
-4	20
-3	769
-2	3237
-1	1856
1	2343
2	4815
3	2842
4	192
5	3

Finally, Graph Score Frequencies. The output should be as follows: **(10 points)**



Task 4: Calculate average sentiment scores for each section of 100 lines. **(15 points)**

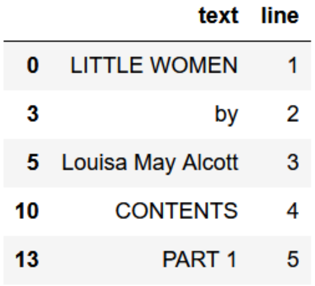
Just divide the text into sections of 100 lines and calculate a sentiment score for each section. The output should be as follows if show the first 10 rows:



Out[9]:

	score
section	
0	0.020408
1	0.362745
2	0.619565
3	0.500000
4	0.871795
5	0.823529
6	0.512821
7	1.082353
8	0.041667
9	0.816092

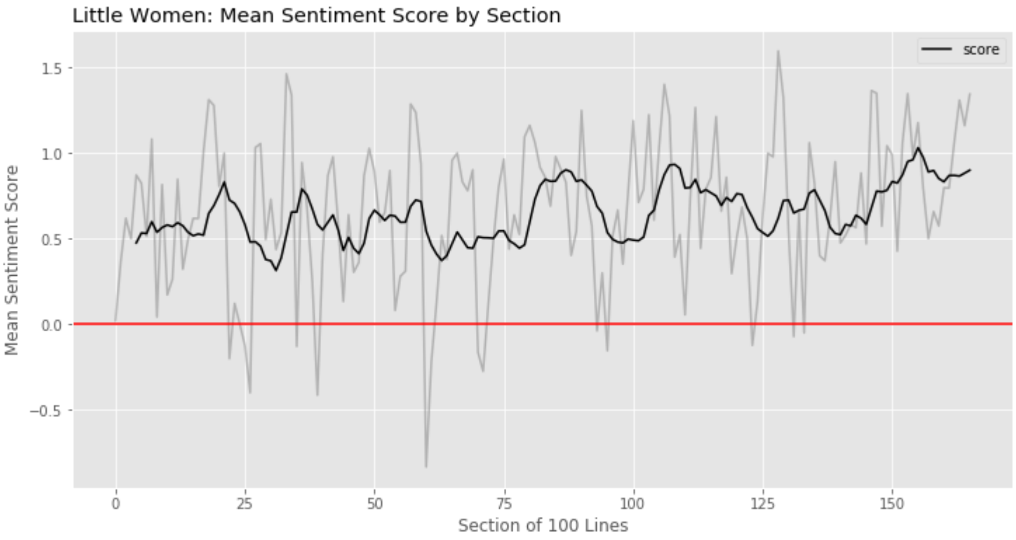
Note that you need add line numbers first. These numbers will be used to divide the text into sections. For an example look at the following output:



	text	line
0	LITTLE WOMEN	1
3	by	2
5	Louisa May Alcott	3
10	CONTENTS	4
13	PART 1	5

Task 5: Graph the "sentiment arc" of the story. **(15 points)**

Plot Scores by Section to View Narrative Arc. The output should be as follows:



The submission grading rubric is as follows (points out of 75 total):

Project element	**Points**
Task 1	10
Task 2	15
Task 3	20
Task 4	15
Task 5	15

Submission Instructions: Create a compressed file (.zip or .tar.gz files are accepted) with your all source files such as .ipynb files and data files. Generally speaking to complete Task1 through Task5, you just need one .ipynb file. But it’s better to submit everything as a compressed file. Submit the compressed file to Blackboard.

Late submission policy: As described in the syllabus, any late submission will the penalized with 10% off after each 24 hours late. For example, an assignment worth 100 points turned in 2 days late will receive a 20 point penalty. Assignments turned in 5 or more days after the due date will receive a grade of 0.