16.10. Project - Sentiment Classifier

We have provided some synthetic (fake, semi-randomly generated) twitter data in a csv file named project_twitter_data.csv which has the text of a tweet, the number of retweets of that tweet, and the number of replies to that tweet. We have also words that express positive sentiment and negative sentiment, in the files *positive_words.txt* and *negative_words.txt*.

Your task is to build a sentiment classifier, which will detect how positive or negative each tweet is. You will create a csv file, which contains columns for the Number of Retweets, Number of Replies, Positive Score (which is how many happy words are in the tweet), Negative Score (which is how many angry words are in the tweet), and the Net Score for each tweet. At the end, you upload the csv file to Excel or Google Sheets, and produce a graph of the Net Score vs Number of Retweets.

To start, define a function called `strip_punctuation` which takes one parameter, a string which represents a word, and removes characters considered punctuation from everywhere in the word. (Hint: remember the *.replace()* method for strings.)

Save & Run Load History Show CodeLens

```
punctuation_chars = ["'", '"', ",", ".", "!", ":", ";", '#', '@']
```

Activity: 16.10.1 ActiveCode (assess_ac_18_1_1_1)

Next, copy in your strip_punctuation function and define a function called `get_pos` which takes one parameter, a string which represents one or more sentences, and calculates how many words in the string are considered positive words. Use the list, `positive_words` to determine what words will count as positive. The function should return a positive integer - how many occurrences there are of positive words in the text. Note that all of the words in `positive_words` are lower cased, so you'll need to convert all the words in the input string to lower case as well.

Save & Run Load History

```

punctuation_chars = ["'", '"', ",", ".", "!", ":", ";", '#', '@']
# list of positive words to use
positive_words = []
with open("positive_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            positive_words.append(lin.strip())
```

Activity: 16.10.2 ActiveCode (assess_ac_18_1_1_2)

Next, copy in your strip_punctuation function and define a function called `get_neg` which takes one parameter, a string which represents one or more sentences, and calculates how many words in the string are considered negative words. Use the list, `negative_words` to determine what words will count as negative. The function should return a positive integer - how many occurrences there are of negative words in the text. Note that all of the words in `negative_words` are lower cased, so you'll need to convert all the words in the input string to lower case as well.

Save & Run Load History

```

punctuation_chars = ["'", '"', ",", ".", "!", ":", ";", '#', '@']

negative_words = []
with open("negative_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            negative_words.append(lin.strip())


```

Activity: 16.10.3 ActiveCode (assess_ac_18_1_1_3)

Finally, copy in your previous functions and write code that opens the file `project_twitter_data.csv` which has the fake generated twitter data (the text of a tweet, the number of retweets of that tweet, and the number of replies to that tweet). Your task is to build a sentiment classifier, which will detect how positive or negative each tweet is. Copy the code from the code windows above, and put that in the top of this code window. Now, you will write code to create a csv file called `resulting_data.csv` , which contains the Number of Retweets, Number of Replies, Positive Score (which is how many happy words are in the tweet), Negative Score (which is how many angry words are in the tweet), and the Net Score (how positive or negative the text is overall) for each tweet. The file should have those headers in that order. Remember that there is another component to this project. You will upload the csv file to Excel or Google Sheets and produce a graph of the Net Score vs Number of Retweets. Check Coursera for that portion of the assignment, if you're accessing this textbook from Coursera.

Save & Run Load History

```

punctuation_chars = ["'", '"', ",", ".", "!", ":", ";", '#', '@']
# lists of words to use
positive_words = []
with open("positive_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            positive_words.append(lin.strip())


negative_words = []
with open("negative_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            negative_words.append(lin.strip())
```

Activity: 16.10.4 ActiveCode (assess_ac_18_1_1_4)

Data file: `project_twitter_data.csv`

```
tweet_text,retweet_count,reply_count
@twitteruser: On now - @Fusion scores first points #FirstFinals @overwatchleague @umich @um
BUNCH of things about crisis respons… available July 8th… scholarship focuses on improving 
FREE ice cream with these local area deals: chance to pitch yourself to hundreds of employe
AWAY from the Internet of Things attacks… right and the left? See… use DIY language to disc
IN City Name!… from @twitteruser has some amazing datasets and tools for collective action.
ENTREPRENEURSHIP in #City.… a great social media job opportunity for our UMSI alumni lookin
BRINGING #datascience to community researchers with a team of researchers have been helping
WHY not pay you?… endure crushing pressures and frigid temperatures but the most difficult 
A bunch of things about crisis respons… – but the implications are much bigger t… for some 
@THEEMMYS nomination for Best Lead Actor in a library role focused on Digital Services and 
THIS headline has been inescapable this summer. Now the infectious chart-topper from 'Scorp
OF wine with a shoe? Yes but it ain't pretty. Its First Amendment rights. That it claims wi
HAVE detained six people in history to ride to orbit in private space taxis next year if al
PET Name. She is 1 year old Shiba Inu. When her caregiver is not at home Name likes to have
YOU’RE welcome! He was a mix of many breeds. He is a 2 year old Yellow Lab. When he was a m
Name. He is wild and playful. He likes to chase and play with his stuffed elephant! the Dir
BORDER Terrier puppy. Name is loving and very protective of the people she loves. Name2 is 
REASON they did not rain but they will reign beautifully couldn't asked for a crime 80 year
```

Data file: `positive_words.txt`

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Opinion Lexicon: Positive
;
; This file contains a list of POSITIVE opinion words (or sentiment words).
;
; This file and the papers can all be downloaded from
;    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
;
; If you use this list, please cite one of the following two papers:
;
;   Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."
;       Proceedings of the ACM SIGKDD International Conference on Knowledge
;       Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,
;       Washington, USA,
;   Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing
;       and Comparing Opinions on the Web." Proceedings of the 14th
;       International World Wide Web conference (WWW-2005), May 10-14,
;       2005, Chiba, Japan.
```

Data file: `negative_words.txt`

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;
; Opinion Lexicon: Negative
;
; This file contains a list of NEGATIVE opinion words (or sentiment words).
;
; This file and the papers can all be downloaded from
;    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
;
; If you use this list, please cite one of the following two papers:
;
;   Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."
;       Proceedings of the ACM SIGKDD International Conference on Knowledge
;       Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,
;       Washington, USA,
;   Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing
;       and Comparing Opinions on the Web." Proceedings of the 14th
;       International World Wide Web conference (WWW-2005), May 10-14,
;       2005, Chiba, Japan.
```

You have attempted 5 of 5 activities on this page

(ChapterAssessment.html)

(../NestedData/toctree.html)

✔ Completed. Well Done!

© Copyright 2017 bradleymiller. Created using Runestone (http://runestoneinteractive.org/) 5.5.17.

| Back to top