

Mid Term Project

Due Date: 11.59 pm Nov 22, 2020

Submit a single notebook file (.ipynb) on Quercus

Background

Sentiment Analysis is a branch of Natural Language Processing (NLP) that allows us to determine algorithmically whether a statement or document is “positive” or “negative”.

Sentiment analysis is a technology of increasing importance in the modern society as it allows individuals and organizations to detect trends in public opinion by analyzing social media content. Keeping abreast of socio-political developments is especially important during periods of policy shifts such as election years, when both electoral candidates and companies can benefit from sentiment analysis by making appropriate changes to their campaigning and business strategies respectively.

The purpose of this assignment is to compute the sentiment of text information, in our case tweets posted during the 2016 Canadian elections, and answer the question regarding: *“Can we use Sentiment analysis on Twitter data to get an insight into the American’s political landscape?”*

Learning Objectives

- How to parse and clean data
- How to write and implement algorithms
- How to analyze an algorithm
- How to analyze and display results

Tool Required

- You can use any built-in functions of Python. Besides, you can use anything in these three packages: Numpy, Pandas, Matplotlib. You are not allowed to use any other packages of Python, unless, the question clearly states.
- Data Files
 - **corpus.txt**: corpus containing a set of words and associated sentiment value
 - **stop_words.txt**: file containing a list of all stop words to delete for tweets
 - **us_election_tweets.csv**: a csv file containing tweet data

To Do

- A. Implement functionality to parse and clean a data by applying each of these functions to all tweets:

- 1- Write a function `lower_tweet(tw)` that takes in as input `tw`, a tweet string. Then, return the same string all in lower case (%2).

```
def lower_tweet(tw):
    '''
    (str) -> str
    Input: a string tw (a tweet line)
    Output: lower case string
    >>> lower_tweet("Hello World!")
    'hello world!'
    '''
```

- 2- Write a function `clean_data(tw)` that takes in as input `tw`, a tweet string, cleans it by removing all punctuations and returns the cleaned tweet as output . (The function must have a return statement) (%5).

```
def clean_data(tw):
    '''
    (str) -> str
    Input: a string tw
    Output: a string whose content is that of tw with
    punctuations removed
    >>> clean_data("living the dream.#tommulcair
    instagram.com/p/8up9qepkxw/")
    'living the dream tommulcair instagramcomp8up9qepkxw'
    '''
```

- 3- Write a function `remove_stop_words(tw)` that takes as input `tw`, a tweet string line, and returns the cleaned (stop words removed) version of the tweet as a string. Use the `stop_words.txt` file for this section. Note that before attempting to remove the stop words, all punctuations should be removed from the lower case tweet. (The function must have a return statement.) (%10)

```
def remove_stop_words(tw):
    '''
    (str) -> str
    Input: a string tw
    Output: a string whose content is tw with stop words removed
    >>> remove_stop_words("living the dream.#tommulcair
    instagram.com/p/8up9qepkxw/")
    living dream.#tommulcair
    instagram.com/p/8up9qepkxw/ '
```

- 4- Write a function, `bag_of_words(tw)`, that takes as input a tweet and creates a bag-of-words for it. A bag-of-words is a proper data structure that lists the number of times a word occurs in each tweet (10%). When called on a string: *drink forgotten table drink*, `bag_of_words` should return a proper Python data structure: `'drink': 2, 'forgotten': 1, 'table': 1`

Submission:

Submit a single notebook file (.ipynb) via Quercus with the following naming convention:

lastname_firstname_assignment1.ipynb

Make sure that you comment your code appropriately and describe your algorithms in sufficient detail. Your module should be self contained, i.e., the functions you submit cannot call functions you defined in other Python modules or Python codes.

Note: DO NOT place any print() or input() statements in the functions you submit.