Practical Data Science: Assignment 1

Setup

Welcome to the first assignment for this course! In this assignment, my goal is to make sure you become familiar with how to work on and deliver a piece of code that does a particular task in Python and how assignments work for this class. To that end, before you even begin working on this assignment, I want you to make sure you have run all the notebooks for the first 3 lectures (you will not need the lecture on databases for this particular assignment) and that you have understood all the key concepts we went over in class. In particular, as we are going to use the Twitter API in this assignment, I want you to study the notebook that deals with this from the Data Crawling module (twitter_api.ipynb), because we did not have time to go through this in class. Don't worry, it is very straightforward – but I do want you to create your own Twitter App as described in that document because you will need it to test your code for this assignment. Once you are confident with the all material for the lectures 1 – 3, it's time to dive into the questions for this assignment.

Questions

There are **three** questions in this assignment and you must complete all three.

Question 1 (5 points)

In this first question we want you to create a function called `get_reply_precentage`. This function should take in the following three parameters:

`config`: A config object exactly like the one we use in the Twitter notebook i.e.

```
config = {
    'consumer_key': 'XXXXXXXXXXXXXXXXXXXXXXXXX',
    'consumer_secret':
'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX',
    'access_token':
'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX',
    'access_token_secret':
'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
}
```

`username`: A Twitter screen name
`max_tweets`: The maximum number of tweets to retrieve from this user. This parameter should default to 100

This function should return the percentage to 1 decimal place of the provided Twitter user's most recent tweets (up to max_tweets) that were replies to someone else. Note that one way to see if a tweet is a reply is that reply tweets often have the format RT @<username>: e.g. RT @ruimiguelforte:

Question 2 (5 points)

Project Gutenberg (www.gutenberg.org) is an online repository of books and texts that has a lot of the world's literature. You should visit the website to determine that it has a handy search feature that allows you to look for texts e.g. for a particular author. In this question we want you to use your experience of web scraping that you garnered from the example on Beer Advocate to parse the results of a search of Project Gutenberg's website. In particular we want you to write some Python code that uses the BeautifulSoup module to programmatically **create and then print out** a list of all the books written by renown British author Charles Dickens that are available on Project Gutenberg. Make sure you visit the website to search for Charles Dickens and examine the page that is returned.

Question 3 (10 points)

Note that this question is worth more points than the previous two. Here is a link to the online text of "A Christmas Carol" by Charles Dickens on Project Gutenberg:

http://www.gutenberg.org/cache/epub/30368/pg30368.txt

(Aside: This in itself should give you a hint that "A Christmas Carol" should be one of the entries returned in the answer to the previous question, if you did it correctly...)

Visit the aforementioned link and notice that it contains the complete text of the book as well as some material outside the book pertaining to license information and usage at the very end of the file.

For this exercise, we want you to write some Python code that will visit this website to retrieve this file and print out all the **sentences** that are part of the story that contain the name of the protagonist, Scrooge, at least once. To help you, here are the first two sentences (which you can confirm by visiting the link):

```
Scrooge signed it: and Scrooge's name was good upon
'Change, for anything he chose to put his hand to.

Scrooge knew he was dead?
```

HINT: In case it is not clear, you'll need to use regular expressions in order to look for the word Scrooge and to describe what it means for a sentence in English to contain at least this word once.

Submission Instructions

You must submit your assignment as a Jupyter notebook that will contain the full code along with any relevant documentation. The Jupyter notebook must be fully replicable: that is, somebody reading it must be able to do exactly what you did and obtain the same results. The documentation must be at the level where somebody that has some Python knowledge can understand exactly what you are doing and why. In particular, I want to be sure I know what code is answering which question in your submission. Please do not write long essays or repeat the questions in your notebook. Also, please be sure to submit any text in English not Greek.

Honor Code

You understand that this is an **individual** assignment, and as such you must carry it out alone. You may seek help on the Internet, by Googling or searching in StackOverflow for general questions pertaining to the use of Python and its libraries and idioms. However, it is not right to ask direct questions that relate to the assignment and where people will actually solve your problem by answering them. You may discuss with your fellow students in order to better understand the questions, if they are not clear enough, but you should not ask them to share their answers with you, or to help you by giving specific advice.