**Assignment 2**

Web Analytics ISGB/BYGB 7978

Name\_\_Tyler DiNapoli-Chiappelli\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID\_\_\_**A18265560** \_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Any late submissions will have points deducted as follows: 1 day late: -15%, 2 days late: -30%; not accepted for grading if submitted after 3 or more days. Any such assignments will receive score of zero.
2. Late submission due to special occasions can be discussed separately with the instructor.
3. Please submit a **ZIPPED FILE** with the following: word file with your answer, your code files, and the CSV files to the Blackboard. The Assignment 2 is Due on 10/19/2020.

**Problem 1: Twitter Crawling**

Web crawling technique helps companies to analyze their new product launch response and benchmark against competitors. You will explore how Twitter can be used for social media listening in this homework.

Please see the submitted Python program for the code and please see the Wordclouds below for images, accompanying tables, and explanations.

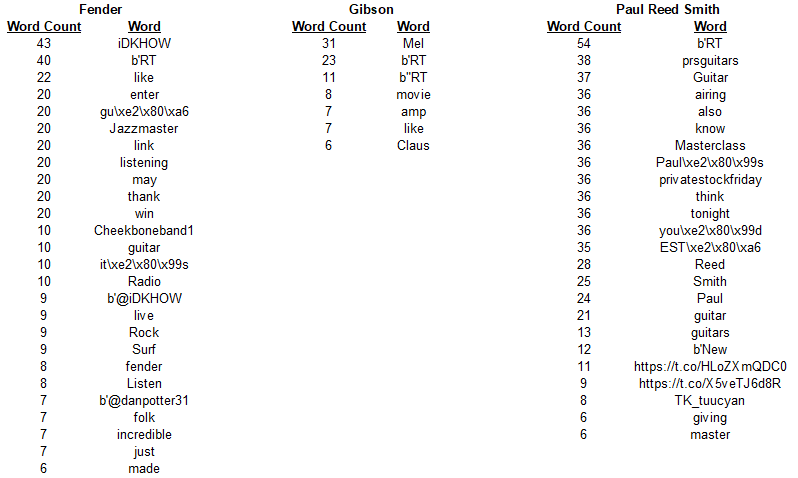
1. Identify a focal product/service of interest. For example, iPhone 11. Identify 2-3 competitor’s products. (This can also be a Website, for example, Amazon vs. eBay and Target, or even celebrities.) Pick 3 product features of interest. For example, if you choose iPhone and Samsung, the comparison feature can be Price, Screen, Size.
2. Prepare a list of hash tags or relevant keywords that are relevant to your focal product and the competitor’ products. For example, “#iPhone11”, “#iPhone11Plus”.
3. Use the Tweepy program introduced in class to crawl tweets for all products. You should get at least 100 tweets for each crawling. If you have multiple hashtags or if you crawl several times, you will get more than 100 tweets.

(Alternatively, you could use Twint to crawl tweets for all products.

1. Process the crawled data and keep the following information in a csv file: user screen name, number of Followers, location, Tweet text. Include this processed file as part of your assignment to be submitted.
2. Use a Word Cloud tool (such as <https://www.wordclouds.com/> or <http://www.wordle.net/create>), paste the crawled Twitter content and generate 3 (or 4) Word Cloud for the focal product and the competitors. An example of a Word Cloud is shown below:



1. Observe the results and evaluate if the three product features you picked appeared in the Word Cloud. Based on your observation, provide a comparison table of the focal product and its benchmarks.



Above are the numerical results for the Wordclouds based off crawling 100 tweets for each of “Gibson”, “Fender”, and “Paul Reed Smith”. As you can see in the submitted CSV files, there are far more results than what is displayed here. However, for the purposes of this table, I opted to truncate each result at Word Count < 6.

Prior to knowing the results, here is a summary of my intentions:

**Focal Product:** Gibson guitars

**Competing Brands with the "focal product":** Fender, Paul Reed Smith

**Product Features of Interest:** finish, pickups, wood type

**Selected Hashtags:** #guitarfinish, #guitarpickups, #guitarwood

As can be seen from the Wordcloud charts, the contents of the hashtags I selected are absent entirely. As a matter of fact, for “Gibson”, likely due to famous actor Mel Gibson, musical words are almost entirely absent, with the exception of “amp”. This is by itself surprising because while Gibson does manufacture amplifiers, they are known almost exclusively for their guitars.

Two of Gibson’s competitors, Fender and Paul Reed Smith, seem to have more musical words associated with them on Twitter, but not by a whole lot. “Guitar” (and it’s related forms) and “Jazzmaster”, a model of Fender bass guitar, do appear in these lists with a high word count but are still quite outnumbered by an assortment of other terms.

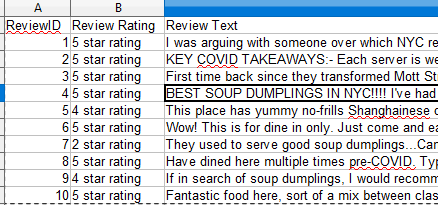
**Problem 2: Yelp Crawling**

BeautifulSoup is a useful tool to crawling structured web content such as Yelp, Amazon, TripAdvisor. In this assignment, you are requested to crawl information about a restaurant near a zip code of your choice from Yelp website. (Please notice that we only want the actual results, not the advertisements).

1. Identify the URL of the ice cream shop of interest.
2. Write a program using BeautifulSoup to crawl the following information:
3. The first 100 reviews of the restaurant. If the business has less than 100 reviews, crawl as many reviews as you can.
4. For each review, please also identify the star rating associated with the review.
5. Write your final result into a .csv file, with three columns: reviewID, reviewRating, reviewText.
6. Submit the csv file and your Python program.
7. In the word document, please include the URL of the restaurant and the first 10 records in the CSV files.

**Restaurant URL**: https://www.yelp.com/biz/shanghai-21-new-york-2

**First 10 csv records screenshot (the review takes up too much space):**

****

