**Analytics for Unstructured Data: Group Assignment #1**

**New due date: 09/21 by 23:59 hours on Canvas**

In this assignment you have been hired as an analytics consultant by JD power and Associates, who wants to perform a competitive analysis of the entry level luxury car market in the USA. Your job is to give advice/insights to these individuals based on the analysis of social media conversations. The detailed tasks are described below.

1. Write a scraper to fetch messages posted in Edmunds.com discussion forums. The scraper output should be a .csv file with the following columns: date and message (even though you will only use the messages in your analysis). Before you develop the scraper, carefully study one of the forums on Edmunds.com to understand the html as well as the threading structures.
2. Fetch around 5000 posts about cars from the Entry Level Luxury forum <https://forums.edmunds.com/discussion/2864/general/x/entry-level-luxury-performance-sedans>

The idea is to have multiple brands and models being discussed without one of them being the focal point. You can choose early or recent posts (do mention what you have chosen). Note that Edmunds changed its forum structure a few years ago, but left the early posts with the old structure. So you should choose either the oldest or newest posts.

**Task A:** Once you fetch the data, test if the data support Zipf’s law. Plot the most common 100 words in the data against the theoretical prediction of the law. For this question, do not remove stopwords. Also do not perform stemming or lemmatization.

*Hint: Check* [*http://www.garysieling.com/blog/exploring-zipfs-law-with-python-nltk-scipy-and-matplotlib*](http://www.garysieling.com/blog/exploring-zipfs-law-with-python-nltk-scipy-and-matplotlib)

Task B: Find the top 10 brands from frequency counts. You will need to write a script to count the frequencies of words (stopwords should NOT be counted). Replace frequently occurring car **models** with **brands** so that from now on you have to deal with only brands and not models. You will need another script for this job. A list of model and brand names (not exhaustive) are provided in a separate file.

**Task C:** Calculate lift ratios for associations between the top-10 brands identified in Task A. You will have to write a script to do this task). For lift calculations, **be sure not to count a mention more than once per post, even if it is mentioned multiple times in the post.**

**Task D:** Show the brands on a multi-dimensional scaling (MDS) map (use a Python script for MDS, there are multiple scripts available on GitHub).

**Task E:** What insights can you offer to your client from your analyses in Tasks C and D

There are some noticeable pairs having high lift values. Combined with the result of Multi-Dimensional-Scaling plot, Volkswagen and Audi; Volkswagen and Mercedes-Benz; and Audi and Mercedes-Benz constitute a group. Toyota and Honda, and Toyota and Nissan are the other group.

From the high lift value from a forum, we can estimate that When people talk about Volkswagen, they are likely to mention about some topic related to Benz or Audi. In the same vein, Audi is likely to be co-mentioned with Benz.

The fact that these three brands comprise a group which might be represented by the image of elegance and that Japanese brands lineup also constitutes another group does not come to surprise.

If there is someone who consider buying a car and simultaneously think of convenient re-sell as a priority, we can recommend to buy a car which is contained in one of this two groups.

On the other hands, what attracts our intention is the rather independent position of the BMW. The lift value of BMW with any other car is not greater than 2. Albeit, we can we some association of BMW with Benz through the MDS, still BMW is located alone.

Based on this result, the person who doesn’t want to have a car which is likely to be compared and considered with another car can choose BMW as his first option. However, one should keep in mind that this lift-based association can be linked to the switching behavior, even though some deviation from the lift can happen!

**Task F:** What are 5 most frequently mentioned attributes or features of cars in the discussions? Which attributes are most strongly associated with which of these 5 brands? You DON’T have to do a sentiment analysis for this assignment.

**Task G:** What advice will you give to your client from Task F? For this assignment, you can assume that all sentiments are positive.

The brands most associated with luxury are Kia, Lincoln, Subaru, Hyundai, and Buick. Three of the five brands are foreign brands while Lincoln and Buick are the luxury brands of Ford and GM, respectively. All the brands listed are more commonly known for their middle-class affordability rather than the conventional idea of luxury. This may mean that consumers compare luxury to price.

The brands most associated with price are Suzuki, Volkswagen, Audi, Kia, and Honda. Only Kia is associated with both luxury and price. Noticeably, manufacturers like BMW and Mercedes-Benz are not associated with price. This may mean higher levels of luxury that brands such as BMW are considered diminishing returns by consumers. That the cars are too expensive compared to other brands that are affordable luxury.

For the other top attributes, there is not a lot of cross-over between brands and the different attributes. This is likely due to the specialization of different brands. Different brands focus on differentiating themselves from competition by focusing on certain aspects of the car. This can be seen regarding Chevrolet as they appear as the most associated with handling but do not appear associated with any of the other attributes.

I would advise my client that the competitive landscape of the entry level luxury car market in the USA seems to buck conventional wisdom. The cars most associated with luxury are not the pricey BMW’s or Mercedes-Benz’s but rather the Kia’s. Entry-level consumers may prefer a modest level of prices for luxury cars. However, most brands attempt to specialize in a key attribute to differentiate themselves from competition.

**Task H:** Which is the most **aspirational** brand in your data in terms of people actually wanting to buy or own? Describe your analysis. What are the business implications for this brand?

According to the aspr\_matrix2 output, we understand that Mitsubishi is the most aspirational brand since it has highest lift value associated with list of aspirational bi - gram words. The implication of such high lift value is that among people's comment discussing about different type of car brands is that Mitsubishi appears to be the most aspirational brand. However, such insight only suggests that people want to buy Mitsubishi but not whether they did buy it. Mitsubishi was not a top 5 brand for any of the common attributes that people talked about, such as price, handling, etc. This may indicate that many people wanted to buy a Mitsubishi but that the brand did not meet their expectations.

**Provide the following details in your python notebook:**

1. Which forum you chose (provide URL)
2. Which 10 brands you chose – provide the frequency table
3. Show all lift calculations in a table.
4. MDS map
5. State the 5 attributes you chose (again, a table is good here).
6. For task E, provide all details of your analysis – e.g., how you measured “aspirational” and how you found the most aspirational brand.
7. Advice/insights based on your analysis for your client.

Your submission (python notebook) should include all scripts as well as your answers to the questions above (generally speaking, I won’t run these scripts, but if the numbers don’t look right, I may run some of them). There is no need to include your data file(s). Please write team member names inside the notebook.