Analytics for Unstructured Data: Group Assignment #1

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 design 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 econometrically. Also 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.

Check http://www.garysieling.com/blog/exploring-zipfs-law-with-python-nltk-scipy-and-matplotlib

(Note that the above link does not test Zipf's law econometrically)

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

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.

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?

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 values in a table.
4. MDS map
5. State the 5 attributes you chose (again, a frequency 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). Please write team member names **inside** the notebook and not as a part of the file name.