Social Media Text Analytics - Edmunds

PREREQUISITES

1/. We developed a scraper using Selenium and BeautifulSoup, taking data from the Edmunds car forum platform, focusing on a forum with subject: "Entry level luxury performance sedans", the url is presented for reference purposes:

'https://forums.edmunds.com/discussion/2864/general/x/entry-level-luxury-performance-sedans'

2/. Scraping a total number of 4000 entries we saved the results in a .csv file with dedicated columns for date, userid and message using the structure of the forum's website. The forum chosen was one with entries dedicated to multiple brands and models in order to remove the data bias from our analysis.

3/. Initiating our analysis, we found the companies with the maximum number of mentions in the collected list of messages. In order to ensure the integrity and accuracy of the data, we used a list of car models and replaced the name of the models that appeared and replaced them with the corresponding brands. At the same time, the number of mentions of a brand within the same message were not counted more than once, in order to reduce the effect of the car model changing name (e.g. in case a post mentioned both the brand and the model it would result in two mentions while only one would be accurate). Based on the above, the following were the top 10 most frequently mentioned brands:

```
Out[102]: [('bmw', 2107),
           ('acura', 939),
           ('audi', 683),
           ('infiniti', 649),
           ('honda', 597),
           ('nissan', 485),
           ('seat', 474),
           ('toyota', 403),
           ('ford', 375),
           ('mercedes', 316)]
```

TASK A

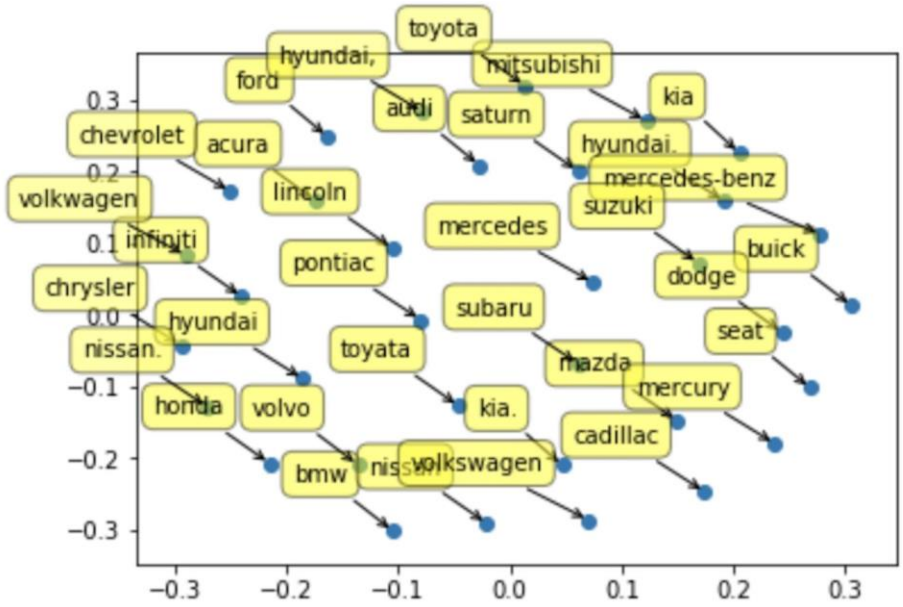
Based on the collected data and identified top 10 brands (BMW, Acura, Audi, Infiniti, Honda, Nissan, Seat, Toyota, Ford, Mercedes - it shall be noted that the brand "Seat" might be an outlier, since people might also be referring to the seats of the car) we calculated the lift values of the

brands focusing on the number of messages that two brand names appear together and came up with the below results:

Lift Table:

	acura	audi	bmw	buick	cadillac	chevrolet	chrysler	dodge	ford	honda	...	pontiac	saturn	seat	subaru	suzuki	toyata	toyota	volkswage
acura	0	9.09091	12.5	1	1	1	1	1	100	4.7619	...	1	1	100	1	1	1	20	
audi	9.09091	0	6.25	1	1	1	1	1	50	50	...	1	1	25	1	1	1	1	
bmw	12.5	6.25	0	1	1	1	1	1	33.3333	33.3333	...	1	1	25	1	1	1	16.6667	
buick	1	1	1	0	1	1	1	1	1	1	...	1	1	1	1	1	1	1	
cadillac	1	1	1	1	0	1	1	1	1	1	...	1	1	1	1	1	1	1	
chevrolet	1	1	1	1	1	0	1	1	1	1	...	1	1	1	1	1	1	1	
chrysler	1	1	1	1	1	1	0	1	1	1	...	1	1	1	1	1	1	1	
dodge	1	1	1	1	1	1	1	0	1	1	...	1	1	1	1	1	1	1	
ford	100	50	33.3333	1	1	1	1	1	0	14.2857	...	1	1	50	1	1	1	9.09091	
honda	4.7619	50	33.3333	1	1	1	1	1	14.2857	0	...	1	1	1	1	1	1	0.909091	
hyundai	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1	1	1	

The results of the brands are also presented in the below multidimensional map (MDS).



The above map is somewhat ambiguous, and more refinement is needed to optimize our results and possibly a larger dataset. In addition, there is room for improvement in regard to the company categorization.

TASK B:

Based on our collected data, we came up with the following list of the top 10 correlated companies.

```
,(('nissan', 'honda'), 1.11)]
,(('toyata', 'honda'), 1.02)
,(('acura', 'infiniti'), 0.95)
,(('honda', 'toyata'), 0.61)
,(('toyata', 'nissan'), 0.56)
,(('bmw', 'audi'), 0.5)
,(('infiniti', 'bmw'), 0.49)
,(('honda', 'acura'), 0.46)
,(('infiniti', 'audi'), 0.37)
[(('toyata', 'ford'), 0.37)
```

Based on our collected data, we came up with the following list of the lowest 10 correlated companies.

```
,(('seat', 'acura'), 0.02)]
,(('acura', 'ford'), 0.03)
,(('seat', 'infiniti'), 0.03
,(('acura', 'nissan'), 0.03)
,(('subaru', 'bmw'), 0.04)
,(('acura', 'seat'), 0.04)
,(('subaru', 'seat'), 0.05)
,(('seat', 'bmw'), 0.05)
,(('honda', 'seat'), 0.05)
[(('subaru', 'honda'), 0.05)
```

As per the above list of brand pairs and corresponding lift values:

- Top three brands are Japanese affordable car manufacturers
- The strongest relation appears between Honda and Nissan, and Toyota and Honda which makes sense since they both have similar branding and marketing models.
- All the lift values appear to be very small with most posts focusing on only one brand which might suggest that customers are focusing on brands separately
- BMW appears relatively high in the list of lifts and it is related to both Audi and Infiniti, but it shall be noted that the data was biased towards BMW since there were many models in the model file that resulted in the stronger brand appearance
- There is a very strong association between Asian brands i.e. people compare an Asian brand only with another Asian brand. However this is not true for American or German counterparts.
- Seat and Subaru are brands which are not often spoken in relation to other brands

We can conclude the following:

1/. Toyota and Nissan are the immediate competitors to Honda. For Honda to strengthen its market share they can expand their strategy focusing on a new strategy in order to create a competitive advantage over Nissan and Toyota and differentiate themselves.

2/. Toyota and Honda are getting compared to every brand. This implies that they are an alternative to every brand which is out there in the market.

3/. In regard to Mercedes, which is a traditional luxury brand, identifying its small correlation to over brands, we would suggest that they have a special place in the market (as it was within the top ten most discussed brands). They should identify the reasons that make them "special" compared to their competition and built up on their competitive advantage.

4/. Honda and Subaru are brands with similar products in regard to price range and quality, but according to customers they present a very different solution and experience. Based on that we would advise Honda not to focus their strategy towards Subaru.

TASK C:

5 attributes we chose:

In order to get the top 5 attributes being discussed about cars, we first calculated the frequency of all words and filtered down a list of features from those words. Next, we created a csv file (output.csv) of replacement words for each attribute that may be described by different words. After iterating through all the messages, we implemented the replacement words and performed a frequency count of the attributes across all messages.

Creating a list of attributes, we came up with the following top 5 categorical attributes as the most important ones:

```
Out[120]: [('performance', 1385),
           ('make', 1343),
           ('condition', 1134),
           ('type', 985),
           ('price', 941)]
```

Performance: refers to values like speed and acceleration of the car
Make: refers to the appearance and exterior features
Condition: refers to the efficiency and optimization of the operation of the car, whether it is new or old or the model is outdated
Type: refers to the category of the car including family, sports car, etc. Price:
refers to the value for money or cost of the car

TASK D:

Below is the list of the most strongly prevailing attributes and the respective brands:

```
,(('seat', 'type'), 0.77)]
,(('saturn', 'condition'), 0.59)
,(('seat', 'performance'), 0.47)
,(('subaru', 'condition'), 0.46)
,(('cadillac', 'make'), 0.37)
,(('saturn', 'price'), 0.36)
,(('hyundai', 'price'), 0.32)
,(('lincoln', 'type'), 0.31)
,(('chrysler', 'condition'), 0.31)
[(('chevrolet', 'condition'), 0.29)
```

Based on the above, we came up with the following useful insights:
1/. Three American brand car owners (Saturn, Chrysler and Chevrolet) are concerned to the condition of the car.
2/. People talking about Cadillac are involved in discussing specifically about exterior features and aesthetics and less concerned about other attributes like performance.
3/. It is interesting that Lincoln does not have a noticeable lift value when compared to the attribute 'luxury' which we would have expected.

TASK E:
Most Aspirational Brand: Acura

Logic: - For calculating most aspirational brand we have selected the brand which have most positive sentiment associated to it. We have not taken into account lifts or negative sentiments deliberately. For eg. let's say BMW has 1000 positive sentiments (i will explain sentiments later) and 1100 negative sentiments and Audi has 500 positive sentiments and 300 negative sentiments, BMW is more aspirational. Reason being BMW is most talked about in positive sense implies people want it the most, but it has huge amount of negative as well associated to it because of which they defer from buying it in the end - However, it is the most aspirational. Its just like a classic value for money scenario. We all want to buy very expensive stuff but we can get more utility from cheaper products. Then, the message is for management that these are the negative attributes which if fixed will make you a leader. All the other parameters like related brands and related attributes are stored in dictionary or lists and can be fetched easily by entering just one key - which is brand name.

Methodology: - For calculating sentiment we have created 2 list's, one with positive sentiment words like 'good', 'better', 'nice' etc. and the other is filled with negative sentiments words like 'not', 'n't', 'bad' etc. When traversing a message we have stored the brands name. Then we calculate attributes associated to that brand by checking whether it is present in that message. THe list of attributes was created manually by studying the words in the website. Reason for this (which we thought) is the attributes should be associated to words spoken in that region and not english dictionary. Once a brand is encountered in a message, we check the attribute present in the

message. If there is one, we pick 5 words to its right and 5 to its left. Use this list to see if it is a positive or negative association.

	cnt_attr
acura	228
infiniti	199
seat	151
volkswagen	81
subaru	74
hyundai	49
buick	30
chevrolet	26
suzuki	21
mazda	19
kia	14
dodge	13
pontiac	11
hyundai.	8
mitsubishi	7
saturn	3
mercury	3

In relation to the most aspiring brand ‘Acura’ ,almost all features are being talked about:

In [348]:
```
set(brand_attribute_pair_pos['acura'])
```

Out[348]:
```
{'brand',
 'condition',
 'engine',
 'interior',
 'look',
 'luxury',
 'make',
 'market',
 'model',
 'performance',
 'price',
 'reliability',
 'safety',
 'sale',
 'speed',
 'steer',
 'technology',
 'transmission',
 'type',
 'wheel'}
```