Jonathan Gaudaire 260966733

INSY 448 – Text and social media analytics

Professor Changseung Yoo

Individual assignment 2

Task A:

The data is scraped using Webscraper. The JSON and selector graph are provided below:

JSON: {"\_id":"yelp\_thai\_food","startUrl":["https://www.yelp.ca/search?find\_desc=thai+food&find\_loc=Montreal%2C+QC&start=[0-120:10]"],"selectors":[{"id":"restaurant\_link","parentSelectors":["\_root"],"type":"SelectorLink","selector":"li:nth-of-type(n+7) a.css-1m051bw","multiple":true},{"id":"boxes","parentSelectors":["restaurant\_link","next\_page"],"type":"SelectorElement","selector":"div.review\_\_09f24\_\_oHr9V","multiple":true},{"id":"next\_page","parentSelectors":["restaurant\_link","next\_page"],"paginationType":"auto","selector":"span.icon--24-chevron-right-v2.navigation-button-icon\_\_09f24\_\_Bmrde","type":"SelectorPagination"},{"id":"review","parentSelectors":["boxes"],"type":"SelectorText","selector":"span.raw\_\_09f24\_\_T4Ezm","multiple":false,"regex":""},{"id":"rating","parentSelectors":["boxes"],"type":"SelectorElementAttribute","selector":"div.five-stars\_\_09f24\_\_mBKym","multiple":false,"extractAttribute":"aria-label"}]}

Selector graph:

![Diagram

Description automatically generated]()

Task B:

After conducting a word frequency analysis, I found that the most frequently mentioned attributes are noodles (1692), service (1557), soup (1284), and beef (1068). As these attributes are almost not replaceable by another word in a review, we only need to replace “noodle”, which appears 825 times, with “noodles”. These attributes were chosen because they could all be included in a search query when one is looking for a restaurant. I also deleted the duplicate reviews, which in the end leaves 3802 reviews to work with.

Task C:

Using the yelp\_reviews dataset with 3802 reviews of Thai restaurants in Montreal with more than 30 reviews, I ran the DocSim script to compute the similarity with the attributes mentioned above, keeping only the 500 reviews with the highest similarity score.

Task D:

I computed the compound sentiment score for the 500 reviews using the sentiment script. I added it as another table column with the restaurant name, review, rating, and similarity score.

Task E:

I first need to compute each restaurant's average sentiment and similarity score to recommend ten restaurants to the user. Using a PivotTable, I find the averages for these two metrics for each restaurant, and I add the average of both scores to rank the 41 restaurants in the dataset. As seen in Appendix 1, some restaurants only have one review among the 500 reviews with the highest similarity score.

The ten restaurants recommended to the user with “noodles”, “beef”, “service”, and “soup” as the four attributes for the recommendations are:

|  |  |  |
| --- | --- | --- |
| Restaurant name | Average similarity score | Average sentiment score |
| Emeraude Bangkok | 0.155 | 0.943 |
| Nhậu Bar | 0.107 | 0.964 |
| Sukho Thai | 0.135 | 0.932 |
| Sen Vang | 0.156 | 0.909 |
| I Am Pho | 0.167 | 0.885 |
| Hà | 0.136 | 0.913 |
| Cuisine AuntDai | 0.129 | 0.917 |
| Satay Brothers | 0.163 | 0.881 |
| Tampopo | 0.189 | 0.853 |
| Bao Bao Dim Sum | 0.107 | 0.927 |

Task F:

The ten restaurants with the highest rating in the entire dataset are:

|  |  |
| --- | --- |
| Restaurant name | Average rating |
| Jardin des Délices | 5 |
| Thai Plus Leon | 5 |
| Malay Thaï | 5 |
| Fine Cuisine | 5 |
| Fusion Express | 5 |
| Krapow | 4.67 |
| Mae Sri | 4.67 |
| Mandy’s | 4.61 |
| Sen Vang | 4.54 |
| The Dumpling Hut | 4.51 |

As we can see, only one restaurant (“Sen Vang”) among the ten highest rated in the dataset is present in our recommendation system.

The first obvious explanation is that most of these restaurants were not included in the similarity score calculation as they had less than 30 reviews. Indeed, only 3 of the restaurants displayed above have more than 30 reviews. Therefore, they could not and should not be included in the recommendations.

Using only the restaurants with more than 30 reviews, the ten highest-rated restaurants are:

|  |  |
| --- | --- |
| Restaurant name | Average rating |
| Mandy’s | 4.58 |
| Sen Vang | 4.57 |
| The Dumpling Hut | 4.48 |
| Cuisine AuntDai | 4.45 |
| Hello 123 | 4.44 |
| Lola Rosa | 4.36 |
| Nouilles de Lan Zhou | 4.33 |
| Pumpui | 4.29 |
| Satay Brothers | 4.19 |
| Wing Phat | 4.17 |

Here again, only three restaurants are included in both tables: Satay Brothers, Cuisine AuntDai, and Sen Vang. This could be because some restaurants recommended only have one or a few reviews with a high similarity score. Therefore, the average scores for similarity and sentiment are only based on one review.

Suppose we limit the recommendation to the restaurants with more fivehan five reviews among the 500 with the highest similarity score. In that case, the ten restaurants recommended are Sen Vang, I Am Pho, Cuisine AuntDai, Café Santropol, Tiradito, Wing Phat, Nouilles de Lan Zhou, Maison Prathet Thai, and Café Saigon. 4 of these are included in the highest rated restaurants in the dataset.

As the rating on a scale from 1 to 5 reflects the reviewer’s sentiment towards the restaurant, it makes sense that the highest-rated restaurants end up being recommended to the user, as the sentiment score is included in the assessment of a restaurant in this model. However, simply recommending the ten highest-rated restaurants would not meet the requirements of the user, as they are looking for good restaurants associated with the attributes “service”, “beef”, “noodles”, and “soup”. Indeed, there are highly rated restaurants in the dataset that probably do not serve noodles or/and soups, for example. Logically, these restaurants should not be included in the recommendations.

Appendix 1: 10 restaurants with the highest recommendation score.

