Insert Your Title Here∗

Insert Subtitle Here

FirstName Surname†  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 [email@email.com](mailto:email@email.com)

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

4 Extraction of Indicative Adjective Phrases

In this section, we import several third-party libraries to do our work. To be specific, we use pandas to deal with the large data set, use spaCy to achieve word tokenization and employ NLTK to extract adjective phrases.

4.1 Data Preparation

Firstly, we load the data set and assign it to an object of type DataFrame which is a kind of data structure of pandas(a 3rd-party library). Then, we extract all business\_id, and choose a business by calling a random number. The business we choose can be seen in the figure below:



Figure 1: business ’s id

Secondly, we collect all reviews of business for extraction of adjective phrases and randomly draw 20000 reviews which will take effect in the step of scoring. All the operations related to data processing are based on pandas.

4.2 Extraction of adjective phrases

Before the process of extracting adjective phrases, we need to tokenize each review and assign a part-of speech to each word in this review. We use spaCy to help us finish those jobs. Then, with the help of nltk, we can get a parse tree from each sentence. By a function named get\_AP\_from\_tree(), we can extract adjective phrase from the parse trees. Specific details are packaged in the function — extract\_adj\_phrase(), which is shown in the figure below:



Figure 2: definition of extract\_adj\_phrase()

As we can see in the above figure, before parsing, we should define the grammar of adjective phrases. After querying the definition of adjective phrases on Wikipedia, we choose several common forms of adjective phrases: “adverb + adjective”, “adjective + preposition + noun” and “adverb + adjective + preposition + noun”. Then, we write them into our code in the form of regular expressions.

After extracting adjective phrase of reviews of business , we count the numbers of occurrences of each adjective phrases, and keep 50 adjective phrases with the most frequencies for subsequent scoring and screening. Those adjective phrases are as follows:

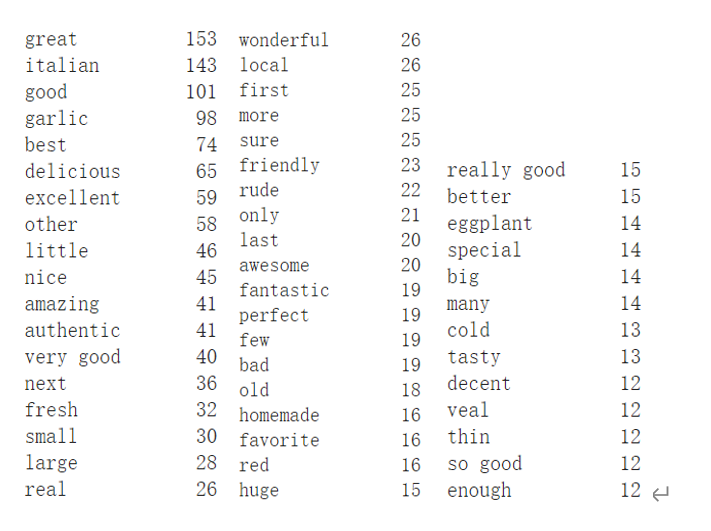


Figure 3: 50 adjective phrases appearing most frequently

4.3 Selection of indicative adjective phrases

To find the indicative phrases which appear often in ’s reviews, but relatively less often in other reviews, we are supposed to apply appropriate scoring method to filter those 50 adjective phrases. The scoring method we choose for each adjective phrases can be indicated in the equation below:

In this equation, is the number of reviews of business , C is the number of all reviews, is the number of times the specified adjective phrase appears in ’s reviews, and w is the number of times the specified adjective phrase appears in all reviews.

Because it takes too much time to count the number of times the specified adjective phrase appears in all reviews, we approximate the effect of all reviews by randomly selecting 20,000 reviews, which means C is the 20000 reviews randomly selected, and w is the times of the specified adjective phrase appears in those 20000 reviews. The code used to calculate the scores of each adjective phrase can be seen in the following figure:

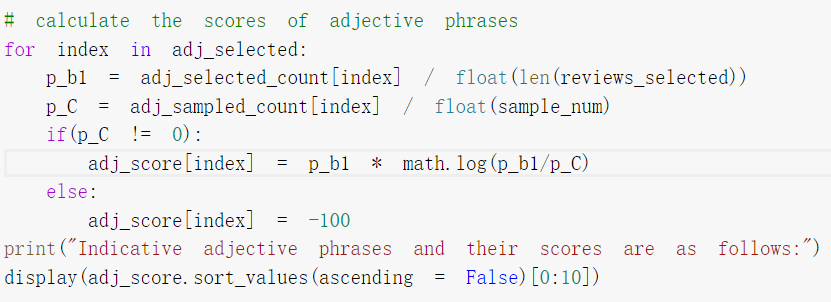


Figure 4: code for scoring

The final result of extracting indicative adjective phrases is shown below:

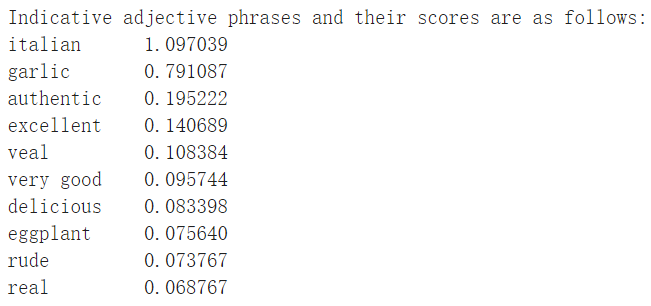


Figure 5: 10 most indicative adjective phrases

From this figure, we can come up with a reasonable guess that the selected business is an Italian restaurant that provide authentic ingredients and delicious food containing garlic and veal.