IST664 Homework1

Wanyue Xiao

1. The data preprocessing section contains dataset importing, data extraction and data saving.

For **dataset importing:** Given that the dataset is consist of lines, one can use readline() to get each line of the dataset and store all these lines to list.

For **data extraction**: In this process, store lines which start with “reviewText:” to a new list variable named “review\_content”. Besides, one can use the strip function to get rid of ‘\n’.

However, instead of using the slicing function to compare data, one can use Regular expression to match line that starts with “reviewText:”. The RE code should be “reviewText(.\*)?\n”.

The **data saving:** In this process, one saves the “review\_content” variable to local directory.

**Code Screenshot:**

**A screenshot of a cell phone

Description automatically generated**

**Result:**

**A close up of a newspaper

Description automatically generated**

1. List the top 50 words by frequency

Firstly, one should tokenize the lines, convert these tokens to lowercase, and use the alpha\_filter function to keep words that only contain alphabetic characters. Since people tends to use common (discriptive) words in reviews, it is better to change contents to lowercase and eliminate numeric values. After executing processes mentioned above, one should also eliminate all the stopwords to ensure the results have true meanings. Finally, given that some of the results have unwanted words with length that is less than two, one can eliminate those words before calculating the world frequency.

**Code Screenshot:**

A screenshot of a social media post

Description automatically generated

**Result:**  
A screenshot of a cell phone

Description automatically generated

From the result, one cannot draw a specific conclusion. According to the word list, it is easy to speculate that these reviews mentioned the size, quality, and color of clothes many times. However, even though there are lots of descriptive words (adj.), one cannot tell the general emotion of those reviewers. Hence, future investigation of ngram analysis should be conducted.

1. List the top 50 bigrams by frequency

The processes of getting the top 50 bigrams is similar to that of getting the top 50 words. After turning the words to lowercase and using the alpha\_filter function to get rid of un-alphabetic characters, one can use the bigrams function to get a list of bigrams. Then, in order to get a list of meaningful results, stopwords should be eliminated from the bigrams list. Finally, FreqDist function will be executed to get the top 50 bigrams by calculating the frequency.

**Code:**

**A screenshot of a cell phone

Description automatically generated**

**Result:**

**A close up of text on a white background

Description automatically generated**

According to the result, customers seem to hold a positive attitude toward the products since “well made”, “would recommend”, and “good quality” appear lots of times in the bigram list.

1. List the top 50 bigrams by pmi

These processes are extremely similar to that of getting the top 50 bigrams by calculating the corresponding frequency. The difference is that the code contains a new filter which rules out words that are below certain frequency. In this case, the threshold is 5. Besides, words with length that is less than 3 have been eliminated. Finally, pmi function has been used to get the top 50 bigrams with corresponding pmi value.

Code:

A screenshot of a cell phone

Description automatically generated

Result:

A close up of a newspaper

Description automatically generated

After analyzing the result, one cannot get specific information from the bigrams list since all of these words are not commonly used in daily life. In this case, one should investigate the meaning of these words before reaching to a conclusion.