**IST 652 Final Project**

**--Amazon product review analysis**

# Data & Source

The project uses Amazon product data collected by Prof. Julian McAuley, an associate professor at UCSD. The data contains about 142.8 million product reviews and metadata from Amazon during May 1996 to July 2014. We mainly focus on the 5-core data of the department of cell phone and accessories.

|  |  |
| --- | --- |
| **Data** | **Description** |
| Product metadata | Descriptions category information, price, brand, and image features |
| Reviews | Ratings, text, helpfulness votes |
| Links | Images of also bought or also viewed products |

# Data preparation

Our data is in good condition and well organized from a trusted resource. It is chosen from a large data file, and we only picked those reviews that users and products have at least 5 reviews.

Data is stored in a json.gz file, so we firstly import module gzip to load the data, and use the gzip.open() function to open the file. The gzip module provides a simple interface to compress and decompress files. After that, we write a loop to pass the data into a dataframe by using pandas module. To take a look at the data and know the basic structure, we return the first five rows of the dataframe.



Here lists the basic introduction of the attributes:

|  |  |
| --- | --- |
| **Attribute Name** | **Definition** |
| reviewerID | ID of the reviewer |
| asin | ID of the product |
| reviewerName | Name of the reviewer |
| helpful | Helpfulness rating of the review. [2,3] means 2/3 |
| reviewText | Text of the review |
| overall | Rating of the review |
| summary | Summary of the review |
| unixReviewTime | Unix time of the review |
| reviewTime | Raw time of the review |

Then, we take out the ‘reviewText’ column from data frame and use to.list() function to convert it into a list of texts. There are 194,439 pieces of reviews in the list.

Before analysis, we use string module to remove punctuations in the list as they are not necessary for our analysis.

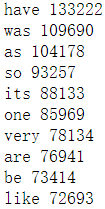
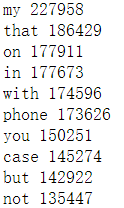
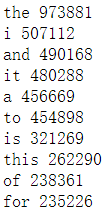
# Analysis

1. Methods
   * NLTK

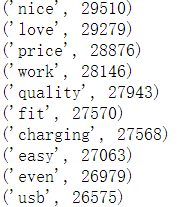
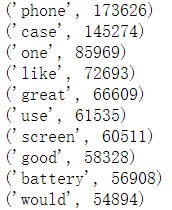
Since we are deal with review data in human language, we need to use natural language toolkit which is also introduced in our advanced topic presentation.

* + Tokenization

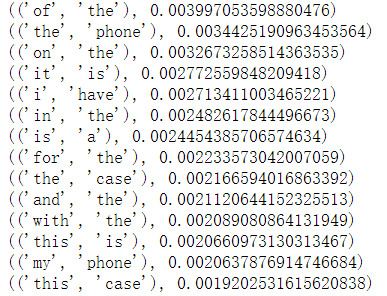
Before next step, we need to tokenize the data in the list. There is a word\_tokenize() function in NLTK, it divides texts into tokens which is useful for sentimental analysis. At the same time, we use FreqDist() function to list out the top 50 tokens in the review.



From these images we could notice that there are some stop words that are not useful for our analysis, so we need to get rid of them. Below are the keywords that customers care about when they choose cellphones and accessories.



These keywords and frequencies cannot tell us the details, so we create bigrams to put two adjacent tokens together to make more sense. We output the top 50 bigrams and their frequencies to see if we could get some insights.

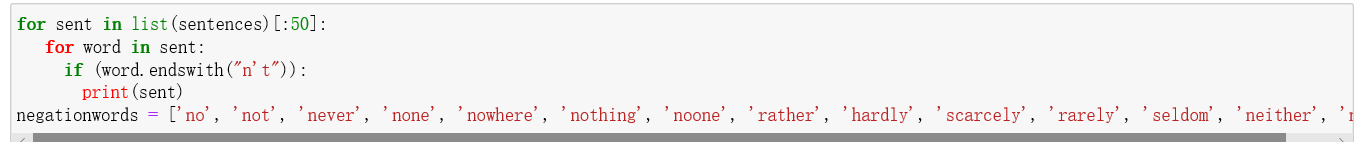


Until then, we noticed that the bigram analysis does not return any useful insights for us. At this time, we should work on the sentiment analysis.

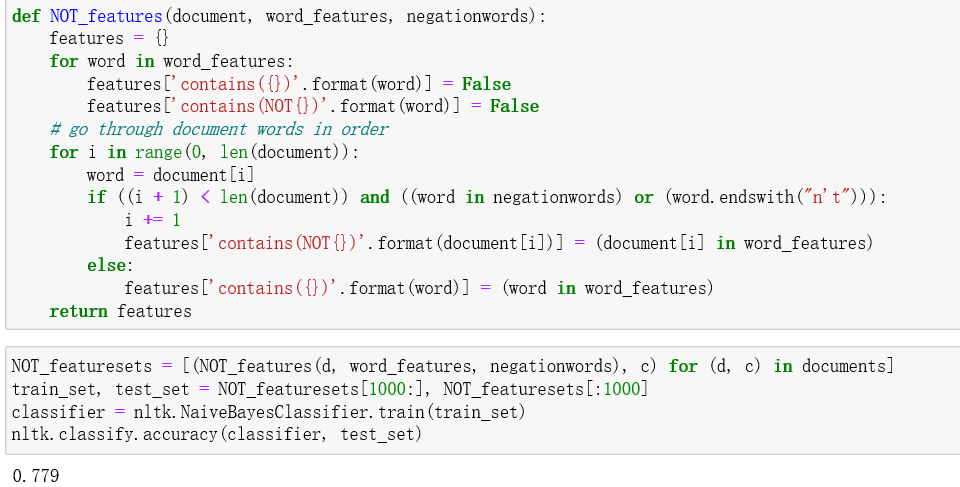
* + Sentimental analysis

In previous work, the results show that the word tokens are not helpful, so we should try on sentences. First step is to tokenize the sentences, after that, tokenize the words in those sentences. For example,” I love IST652! Python is great!”, it would be divided into “I love IST652!” and “Python is great!”. This is more helpful to see if customers are satisfied with certain aspects of the product.

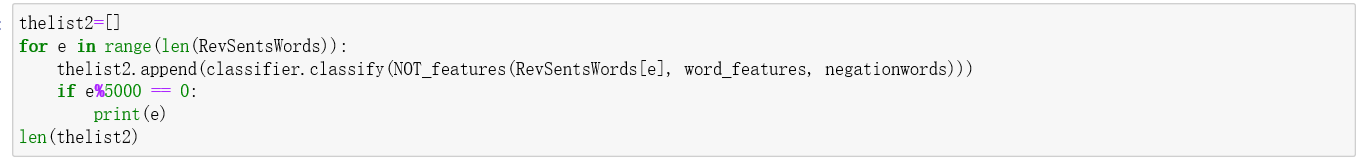
Then, we need sentence\_polarity module from nltk to find out the polarity of sentences which means it could analyze if a word or a sentence is positive or negative.



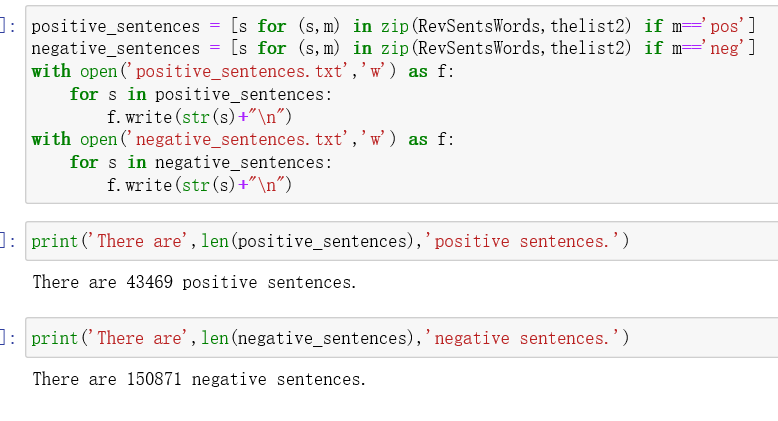
For eliminate the influences from the ‘not’ words, we make the list of negation words.



The Feature sets we make have a 77.9% accuracy which is good for analyzing.



In the long time analyzing period, we made a counter which shows the processing progress.



Finally, we got the sentences and saved them into txt files

We got 43469 positive sentences and 150871 negative sentences out of 194340 sentences of review texts.

1. Questions to be answered
   * What are the most frequently used words in reviews?
   * What are the aspects customer cares about when choosing cellphones and accessories?
   * Are the reviews positive or negative?
   * What are the positive reviews about the products?
   * What are the negative reviews about the products?

# Overall description

This is a group work, Jinwen did the data preparations and the report, and Zequn worked on the analysis. In this project, we mainly focus on the analysis of text review by using NLTK. The first method we use is bigram. After getting rid of stop words, punctuations, and tokenize the data, we find out the frequently used words in reviews related to cell phones. However, the keywords and frequencies are not enough for analysis, we want more detailed information in those reviews. Then we created bigrams to see if paired keywords could give us some insights. We know that customers are talking about the cellphone battery, and customers would like to buy a screen protector and phone case together. After this, we tried sentimental analysis to find out the positive and negative sentences in reviews. Through this step, we could find out the attitude of customers in reviews as a whole. We would know what aspect customers are satisfied with and vice versa.

# Documentation of the output

Besides output in the notebook, there are results in “.txt” files that store the negative sentences and positive sentences separately.

# Conclusions from results

1. “Case” the most frequent word customers mentioned in reviews with 145,274 times.
2. Customers care about the screen and battery when they choose cell phones. They mentioned “Screen” and “battery” in reviews with 60,511 times and 56,908 times.
3. About 22.37% of reviews are recognized as positive whereas the negative reviews take up about 77.63% of the reviews. We can suggest that cellphone business owners need to make some improvements as more than a half reviews are negative.