Text Analysis Report

Silu Zhao

Problem:

1) Scrape at least 5000 reviews/blogs from the web using Python. Stem both lexicon and reviews using NLTK. Find a suitable scraping tools in Python. Prepare a labeled dataset with positive and negative labels in two different ways. The first is by applying the sentiment analysis technique that you have already learnt using a downloaded positive/negative lexicon (e.g. Bing lexicon). The second is by scraping associated user ratings along with reviews. Then a review with four or more stars is labeled as positive and a review with one or two stars is labeled as negative.

2) Divide the labeled dataset into training and test datasets (e.g. 80%/20%). Run NBC in Python on the labeled training dataset for sentiment classification and then show the performance using the test dataset. Your input corresponding to a review is its sparse representation.

Analysis:

In this Problem, I choose the reviews on best buy Samsung s7 glass protector. (https://www.bestbuy.com/site/reviews/zagg-invisibleshield-glass-screen-protector-for-samsung-galaxy-s7/4953402?page=)

There are 5614 reviews from review page 1 to 200.

   1> Scrape

   There are three scraping files. "scrape_bestbuy" is to scrape the text reviews. "scrape_rating" is to scrape the rating numbers. "scrape_rating_and_review" is to scrape the reviews as well as ratings.

   Scrape is the based on the html tag. Reviews are in <li class="review-item"><p class ="pre-white-space"></p> and ratings are in <li class="review-item">. So we get the content inside the parent-child tag with the same class name, which is the rating and reviews.

   The program results in two files:

a. Scape1.csv. It contains only the review of the product

Samsung threw a curve with the standard Galaxy S7| Like the S6 the S7 looks flat| but unlike the S6 the S7 screen has is a slight curve down on the edges| Zagg
Bought both the s7 and the Zagg InvisiSheild Glass| Actually| I bought two of the Zagg products| but| not because I wanted to| Let me explain|| I bought the pho
I used to go commando| no screen protection and no case|||until I dropped my (fortunately very inexpensive and old Motorola) phone 2ft onto the street as I was
I made the big mistake of not reading the reviews for this before I bought it| I've purchased Zagg products before and have always been happy|||especially with
The Zagg glass covering my (regular) Samsung S7 is terrific - with one major drawback| It seems that even my "regular" S7 has a slight curve in the screen on t
Not the best fit around the edges as others have stated| The S7 (non edge version) does have a slight bend around the edge of the glass| So the zaag glass scre
Even after being assured by Best Buy rep (who was great BTW & did look at these reviews with me) that this thing would fit| it doesn't| Not even a week later &
I have always used| recommended| and even sold Zagg screen protectors (as a wireless professional) but I must admit| these S7 "Invisible Shield Glass" suck!!!
I have used Zagg screen protectors for all my devices| and if I need to replace my crack screen protector| not a problem when I can just pay $5 for a replaceme
Let me say that I love Samsung and the Galaxy phones as well as Zagg| in the past I have had nothing but great experiences with both products and companies
I had Zagg glass piece on my S5 and then I threw my phone out of anger and it shattered the Zagg glass but kept my S5 screen intact| I called Zagg n they sent
I had a different brand screen protector from verizon store for 2 years on my HTC One M8 and it worked great| I upgrade to the S7 and got the zagg screen pro
When I originally bought this product it was $40| Immediately after putting it on| the screen protector did not adhere to my phone on the outside edges at all| I h
Bought this screen protector for my Galaxy S7 and was not impressed with how it looked on my phone| It did fit onto my screen but all around the edge of it| it l
I wanted to clear up the complaints about the edge of the screen protector having a haze or not making complete contact| It is true| however

b. Scape3.csv. It contains reviews and rating of the product

2	Samsung threw a curve with the standard Galaxy S7| Like the
1	Bought both the s7 and the Zagg InvisiSheild Glass| Actually| I bou
1	I used to go commando| no screen protection and no case|||until I
1	I made the big mistake of not reading the reviews for this before I I

2> Sentiment Analysis

Sentiment Analysis based on Bing Lexicon. It reads the positive word list and negative word list as reference. Then it tokenizes the reviews, compares with the word lists and counts on the number of positive words and negative words.

If number of positive words > negative words, label it as positive. Mark it as negative otherwise.

	label
on the edges Zagg has established a great reputation for producing so	positive
bought the phone and screen protector though the screen protector v	positive
e street as I was getting out of my car I cant tell you how many times	positive
yespecially with their screen protectors I never purchased the glass p	positive
he screen on the long edges Therefore the last 18 or so of the sides o	negative
e zaag glass screen protector does not fit well on the outside edge It	positive

3> Stemming

We use the labeled data set, negative, positive word lists to do the stemming. "nltk" is used to do the stemming. Then we get the stemming reviews, negative words and positive word lists.

labeledDataset.csv -> stemmingReview.csv
negative.txt -> negative_words.csv
positive.txt-> positive_words.csv

4> Rating anlysis

It reads the file scape3 as the input, which contains reviews and ratings. Then it read the rating and check if the rating is >= 4, the review is labeled as positive, negative otherwise.

label	rating	review
negtive	2	Samsung threw a curve with the standard Galaxy S7| Like the S6 the S7 looks flat| but unlike the S6 the S7 screen has is a sligh
negtive	1	Bought both the s7 and the Zagg InvisiSheild Glass| Actually| I bought two of the Zagg products| but| not because I wanted to|
negtive	1	I used to go commando| no screen protection and no case|||until I dropped my (fortunately very inexpensive and old Motorola)
negtive	1	I made the big mistake of not reading the reviews for this before I bought it| I've purchased Zagg products before and have alwა
negtive	3	The Zagg glass covering my (regular) Samsung S7 is terrific - with one major drawback| It seems that even my "regular" S7 has
negtive	2	Not the best fit around the edges as others have stated| The S7 (non edge version) does have a slight bend around the edge of
negtive	1	Even after being assured by Best Buy rep (who was great BTW & did look at these reviews with me) that this thing would fit| it d
negtive	1	I have always used| recommended| and even sold Zagg screen protectors (as a wireless professional) but I must admit| these S
positive	5	I have used Zagg screen protectors for all my devices| and if I need to replace my crack screen protector| not a problem when I

5> NBC

Input is the stemmed labeled sentiment analysis result. The data set is divided as 80% as the training data and 20% as the test data.

The accuracy is 0.85 and the top 15 informative features are as follows:

```
the accuracy of Naive Bayes classifier is:0.848297213622291
Most Informative Features
                amaz = True            positi : negati =      7.8 : 1.0
               stuck = True            negati : positi =      7.4 : 1.0
                pain = True            negati : positi =      7.2 : 1.0
                bump = True            negati : positi =      6.4 : 1.0
               secur = True            positi : negati =      6.0 : 1.0
                suck = True            negati : positi =      5.9 : 1.0
            perfectli = True           positi : negati =      5.8 : 1.0
                fell = True            negati : positi =      5.5 : 1.0
              poorli = True            negati : positi =      5.5 : 1.0
               gener = True            negati : positi =      5.5 : 1.0
                junk = True            negati : positi =      5.5 : 1.0
               provid = True           positi : negati =      5.1 : 1.0
                 bad = True            negati : positi =      5.1 : 1.0
               clear = True            positi : negati =      5.0 : 1.0
                hang = True            negati : positi =      4.7 : 1.0
```