

Name: Anne Sai Venkata Naga Saketh

USC ID: 3725520208

USC Email: annes@usc.edu

CSCI 544 – Applied Natural language Processing – Assignment 1

Question 2: Print the average length of the reviews in terms of character length in your dataset before and after cleaning.

=====Printing the Average length of Reviews Before and After Cleaning=====

Average Length of Reviews (Before Cleaning): 318 characters

Average Length of Reviews (After Cleaning): 300 characters

I have used the BeautifulSoup to remove the hyper links in the text, used notna to filter out the reviews that are not text and using the Regular expressions to remove any special characters from the text in the review. Used a dictionary with the list of contractions to convert the contractions in the review text.

Question 3: Print three sample reviews before and after data cleaning + preprocessing.

In the .py file, print the average length of the reviews in terms of character length in before and after preprocessing.

===== Printing Sample Reviews Before and After Pre-processing =====

Sample Review 591443 Before Pre-processing:

printer did not work at all as the carriage was stuck in the far right position

Sample Review 591443 After NLTK Pre-processing:

printer work carriage stuck far right position

Sample Review 1498236 Before Pre-processing:

product was as advertised and is a great teaching tool tool set provides large visuals for students and offers a variety

Sample Review 1498236 After NLTK Pre-processing:

product advertised great teaching tool tool set provides large visuals student offer variety

Sample Review 2161966 Before Pre-processing:

ive had this for about a year i had my nd staples mailmate die on me in years with pretty light home use i bought this amazon basics but have been disappointed with performance its underpowered jams often get stuck running too bad ive liked all the other amazon basics products ive purchased

Sample Review 2161966 After NLTK Pre-processing:

ive year nd staple mailmate die year pretty light home usei bought amazon basic disappointed performa
nce underpowered jam often get stuck runningtoo bad ive liked amazon basic product ive purchased

=====Printing the Average length of Reviews Before and After Pre-processing=====

Average Length of Reviews (Before NLTK Processing): 300 characters

Average Length of Reviews (After NLTK Processing): 190 characters

Used nltk library, and the word tokenize to tokenize the words and lemmatize to group the words with t
he same meaning and then converted the whole review to a lower case.

Question 5: Perceptron: Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of your dataset.

===== Training Set Metrics: (Perceptron) =====

Accuracy: 0.91620625

Precision: 0.9258094215129661

Recall: 0.9049458172409914

F1-score: 0.9152587367502892

===== Testing Set Metrics: (Perceptron) =====

Accuracy: 0.83635

Precision: 0.8197555523850287

Recall: 0.8621517531135897

F1-score: 0.8404193076548025

Built the perceptron model and trained it on the training data and calculated the Accuracy, Precision, Re
call and F1 score for the training data and testing data.

Question 6: SVM: Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of your dataset.

===== Training Set Metrics: (SVM) =====

Accuracy: 0.97399375

Precision: 0.974595149300428

Recall: 0.9733648305773245

F1-score: 0.9739796014082657

===== Testing Set Metrics: (SVM) =====

Accuracy: 0.903925

Precision: 0.8963773807186334

Recall: 0.9133696793877857

F1-score: 0.90479375696767

Built the SVM model and trained it on the training data and calculated the Accuracy, Precision, Recall and F1 score for the training data and testing data.

Question 7: Logistic Regression: Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of your dataset.

===== Training Set Metrics: (Logistic Regression) =====

Accuracy: 0.9125625

Precision: 0.9156665953079548

Recall: 0.9088454760208482

F1-score: 0.912243284949002

===== Testing Set Metrics: (Logistic Regression) =====

Accuracy: 0.8929

Precision: 0.887627695800227

Recall: 0.899614865202821

F1-score: 0.893581081081081

Built the Logistic Regression model and trained it on the training data and calculated the Accuracy, Precision, Recall and F1 score for the training data and testing data.

Question 8: Multinomial Naive Bayes: Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of your dataset.

===== Training Set Metrics: (Multinomial Naive Bayes) =====

Accuracy: 0.88323125

Precision: 0.9019769789454364

Recall: 0.8599372554901447

F1-score: 0.8804555779505391

===== Testing Set Metrics: (Multinomial Naive Bayes) =====

Accuracy: 0.860325

Precision: 0.8706390861376968

Recall: 0.846296203671285

F1-score: 0.8582950769777057

Built the Multinomial Naïve Bayes model and trained it on the training data and calculated the Accuracy, Precision, Recall and F1 score for the training data and testing data.