NLP Assignment2

Sravya Kurra
Sxk180049

Q1.1

Adding Edit tag to the words in square brackets is helpful as we can understand the actual meaning of text. In our model building, we stop adding negation tag if we encounter a word with POS JJR or RBR, which the words with EDIT tag can be. This helps to stop tagging positive words as negative.

If we chose to remove the words inside square bracket, we can be negative tagging the words that are actually not negative. This is a con If we get rid of words inside square brackets.

In case we remove the brackets and consider them as normal words, without tagging them, we would be adding these as features in our training set. And when we normalize this training set, we would be considering the words that the author did not write. This affects the feature distribution.

Q1.2

I chose to use a set to store the negation ending tokens. Because it has to check if the word from our training set is already present in the set already defined. It uses hash table to check this. It's faster.

Q3.1

Precision determines number of correctly classified positive samples that out of all the samples that are predicted positive

Recall determines ratio of correctly classified positive samples to all the samples that are actually positive i.e., that are correctly classified positive samples and wrongly classified negative

F-measure:
This takes weighted average of Precision and Recall and is used when the positives and negatives are unevenly distributed

We need to precision to make sure we have rightly identified the positive sentiment out of all those that are predicted to be positive and recall confirming the rate at which we are identifying the most of the positive sentiment. Finally, F-measure to check if both the rate at which predicted positives and negatives are righteously classified

NLP Assignment2

Sravya Kurra
Sxk180049

Q3.2

Precison:0.6293706293706294
Recall: 0.8157099697885196
F-Measure: 0.7105263157894738

Q3.3
Precision: 0.7746031746031746
Recall: 0.7371601208459214
F-Measure: 0.7554179566563468

Logistic Regression model performs better than that of Gaussian Naïve base model as has higher precision, higher F-measure and acceptable recall value compared to lower precision and lower F-measure and higher value for recall.
This is because we assume that features are independent of each other in Naïve base that is actually not true

Q3.4
top features:
[('too', -3.3533787671768023),
('bad', -2.407413898770099),
('dull', -2.10836018514251),
('still', 1.827029470063836),
('boring', -1.7948485689922575),
('fails', -1.786793976089765),
('best', 1.6002100622026556),
('enjoyable', 1.47850960013485),
('brilliant', 1.414832139032727),
('worth', 1.408384309236949)]

These weights are the coefficients of the features in decision function. We consider absolute value so as to give the details of features that mostly affects the decision of our classifier. Negative value of weight just shows how the feature would vary compared to other features that are being considered.

Q4.1
Precision: 0.8171641791044776
Recall: 0.6616314199395771
F-Measure: 0.7312186978297163

NLP Assignment2

Sravya Kurra
Sxk180049

Top features:
('evaluation_dal', 4.196051009843493)
('too', -3.430796382094066)
('bad', -2.267051086409063)
('dull', -1.9747383729377872)
('still', 1.8406504923862788)
('boring', -1.7668498052175217)
('fails', -1.7053507419932747)
('enjoyable', 1.4703734274544282)
('best', 1.4565588333414339)
('way', 1.3937287627407826)

Higher Precision, comparatively lower recall and lower F-measure. It means we are characterizing many samples as positive i.e., we are predicting more positive sentiment. This can be clearly seen in top features, pleasantness having higher weight, high impact on decision function. Hence positive sentiment is being detected more.

Q5.1
2 complete days
Q5.2
Aishwarya and Manas