Text Classification Report

Jess Nguyen 000747411

# Data Description

The data used was Reuters 21578 which contained 21578 news stories from the Reuters newswire that were collected in 1987. Since its release in 1990, it has become the most widely used corpus for research in text classification. The data was classified by 5 topics "earn","acq", "bop", "veg-oil", and "ship". Training and testing data were split 75% and 25% with 80,918 in training and 26,972 in testing. The split happens every time `classify\_reuter` is called and is given the data set and labels it needs to fit with the type of Naïve Bayes. All 21578 document was used for the data set meaning there was also 21578 labels. With 5 topics, I needed to run 21578 data 5 times equaling to 107,890 labels and 107,890 rows of data.

# Description of the Text Representations

The topic tag was used for all 5 text representation to give it a more consistent result. I had chosen Earn - 3987, acq - 2448, bop - 116, veg-oil - 137, and ship – 305 because each one had a varying number of representation count. After running the get\_labels function to see all the counts.

# Results

|  |  |
| --- | --- |
| COMBINED MATRIX  v1 no word removal w/ bags w/ Multi:  [[24411 948]  [ 219 1392]]  Accuracy: 0.96  Precision: 0.59  Recall: 0.86 | COMBINED MATRIX  v2 no word removal w/ set bags w/ Bernoulli:  [[24492 867]  [ 592 1019]]  Accuracy: 0.95  Precision: 0.54  Recall: 0.63 |
| COMBINED MATRIX  v3 no word removal w/ stemmed bags w/ Multi:  [[24400 959]  [ 222 1389]]  Accuracy: 0.96  Precision: 0.59  Recall: 0.86 | COMBINED MATRIX  v4 no word removal w/ set stemmed bags w/ Bernoulli:  [[24461 898]  [ 605 1006]]  Accuracy: 0.94  Precision: 0.53  Recall: 0.62 |
| COMBINED MATRIX  v5 word removal(stopped words) w/ bags w/ Multi:  [[24432 927]  [ 191 1420]]  Accuracy: 0.96  Precision: 0.61  Recall: 0.88 | COMBINED MATRIX  v6 word removal(stopped words) w/ set bags w/ Bernoulli:  [[24615 744]  [ 362 1249]]  Accuracy: 0.96  Precision: 0.63  Recall: 0.78 |
| COMBINED MATRIX  v7 2-gram w/ bags w/ Multi:  [[24630 729]  [ 292 1319]]  Accuracy: 0.96  Precision: 0.64  Recall: 0.82 | COMBINED MATRIX  v8 2-gram w/ set bags w/ Bernoulli:  [[24667 692]  [ 548 1063]]  Accuracy: 0.95  Precision: 0.61  Recall: 0.66 |

# Discussion

The purpose of these classifications is to ensure the (labels) topics: "earn","acq", "bop", "veg-oil", and "ship" match the body, title, dateline, and text contain. Therefore, the better results are ones with higher percent in recall because there would be less complaints when there’s less false negatives on the news than false positives since you can one only have either precision higher or recall higher. Version 5 with word removal of stopped words, bags of word, and Multinomial distribution has the highest recall with 88 percent. Bagging the words to receive the number of occurrences improves the predication since it ensures the true presence and authenticity of the word in document. Removing the stopped words helps by reducing the number of words it needs to predict on and will bring in focus on the topic words. Using Multinomial distribution is a better fit for this data set since it typically used for text classification and is the step to Bernoulli. Multinomial uses the idea of counting the word occurs in the document similar to bags of word. This is preferred because it the number of occurs directly relates to if the topic matches the document contain.

# Future work

If I had more time I would have like to see different tags being used and seeing the results of tags with low counts. The processing on running 21578 documents was huge therefore I would be interested to see what the results would be like if all 120 topics was done. I couldn’t give tf-idf a try as well since lots of my time was spent waiting for python to process 21578 documents.