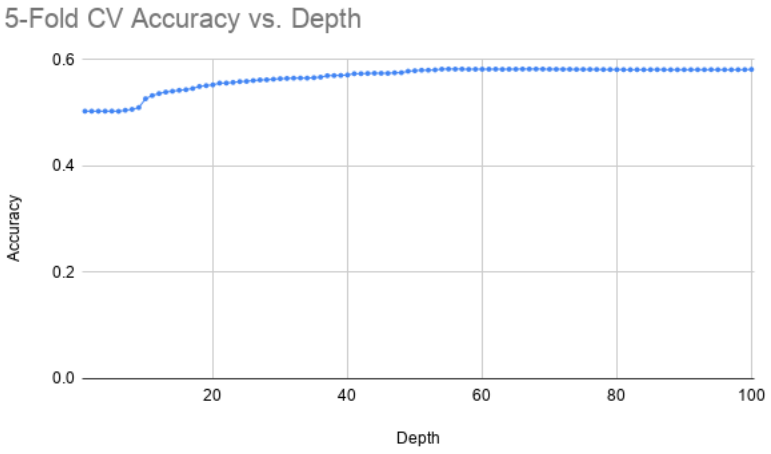
Project Checkpoint 1

Jakob Horvath - u1092049

October 18, 2020

For my first submission to the *Old Bailey* project, I decided to use the *bag-of-words* data set and limited-depth decision tree algorithm. It was quickly discovered how inefficient the algorithm was when having to process 10,000 potential attributes. This was due to the required information gain computation for every single unused attribute when attempting to add a new node to the tree. From this I knew that the data had to trimmed down in some way if I was to continue using this data-algorithm combination.

The resulting solution to this problem was to trim down the accessible attributes to the k most frequent. I ended up setting $k = 250$, and ran 5-fold cross validation on tree depths ranging from 1-100. Even with these restrictions in place, running through the entire cross-validation sequence took many hours to fully compute the results. The graph below shows the cross-validation accuracies for each depth:



The best depth was shown to be 68, with an accuracy of 0.58268. This depth was then used to build the decision tree on the full training set, which was then tested against the *bow.test.libsvm* data set. Attribute 221 was determined to have the highest information gain, and an accuracy of 0.57022 was reported.

Finally, the tree was tested against the *bow.eval.anon.libsvm*, generating a .csv file of the predictions. The public split on Kaggle revealed the accuracy to be 0.58095. Overall, this implementation was shown to be highly unsuccessful in its predictions of the *Old Bailey* court decisions. Moving forward, I plan on trying out two more unique algorithms that will be more time-efficient than a simple decision tree in preparation for the second milestone.