COMP 6714
Project Part 2
Yu Han
Z5219071

Firstly, I used the same method as part 1 to build the idf dictionary of the documents (initiation function in the program). Then I used four features to do the machine learning which are idf score, minimum distance, number of same worlds, and cosine score.

The tf_idf function is used for calculating the idf score of each candidate, which can show the closeness between the candidates and the mention' s document. I used the same function as part 1 to calculate the idf score.

The minDistance function is used for calculating the minimum steps need to change for each candidate to get the mention. For example, the minDistance of "Olympic Park "and "Olympic Game" is 4, because the latter needs to change 4 letters to get the former.

The cosine function can calculate the cosine score between each candidate and the mention, also can get the count of same words between each candidate and the mention. I tried to use the count divide the length of the mention, therefore, the result of the count will be in a small range.

I used these four features to train the data and also use the same features to label the test data. My program might be overfitting in some cases, but I tried to change the maximum depth and the number of boost round to maximum the accuracy in both cases. Finally, I found that set depth as 7 and round as 2100 is the best choice for both cases, which can get 82.04% in case 1 and 53.42% in case 2. However, if I set the depth as 1 and round as 400, then I can get 59% in case 2, but the accuracy for case 1 will be lower. In order to balance both cases, I used the depth as 7 and round as 2100.

I believe that if this program can use the speech tagging of the mentions and candidates, the accuracy will be much higher, because the program will "understand" the sentences instead of only compare the similarity of each letter.