NLP - Exercise 3

Student ID: 211406343

For the classification part, I decided to use CountVectorizer instead of Tfidf. Both functions perform the same thing except that Tfidf normalizes the results. Thus, Tfidf shows information on the more important words and the less important ones. However, for this assignment I decided I will not be using the extra information Tfidf gives, so I decided to use CountVectorizer which appeared to be faster.

I decided to build my own features vector for each chunk based on both grammar and meaning.

Questions:

1. For both features vectors, there was a great gap in recall between `male` tag and `female` one. Male tag had greater recall. On the other hand, the female tag had slightly greater values in precision. These two results combined mean that the classifier could predict male authors better than female ones, that is by tending to classify chunks to male authors in most of the cases which explains the low precision for male tag compared to the precision of female tag.

2. The results of splitting dataset into train and test datasets are much better. That is because k-fold method gives a great weight for outliers in some of its folds. Therefore, when calculating the average of the k classifiers' accuracy, these inaccurate classifiers will eventually affect the final accuracy and will decrease it.

3. The model based on BoW had better predictions. That is because BoW counts all words in the chunk, thus each word influences the result even if it was mentioned very few times. However, when I built my own features vector for each chunk, it did not relate to each word in the chunk' it only checked whether specific words appeared in the chunk or not and how many times they did appear.

4. We should use the meaning of keywords instead of the grammar. That is because both of female and male authors can write properly since we are talking about authors. We are not comparing the writing skills between children and men, so if there is a difference between men and women in writing it will be in the context itself not in grammar.