Assignment 5

This report presents a description and justification of choices made while building a Naïve Bayes model to classify the relation contained in a document. Sentences are classified into the following relations, publisher, director, performer, and characters.

Justification of Design Decision:

The following preprocessing steps were taken

Text formatting:

All words were converted to lowercase, this is to ensure uniformity with words in both the test and train set.

Alphanumeric characters were taken out to improve the efficiency of implementation.

(Code found in the util.py file under the preprocessing function)

Unknown Words:

Words not found in the vocabulary but present in the test set were discarded completely, while words not found in a feature/relation but present in the vocabulary were smoothened to avoid a zero-probability error.

(Code found in the train.py file under the train accuracy function)

Stop words:

The below table shows a variation in output result classification when stop words were discarded and also when they were used.

As seen below, the accuracy of the test set was higher than the train set when stop words were included in the vocabulary. However, when stop words were discarded, the training accuracy was reduced by 0.001 and was higher than the test accuracy.

	Test	Train
With stop words (avg accuracy)	0.893	0.883
Without stop words (avg accuracy)	0.870	0.882

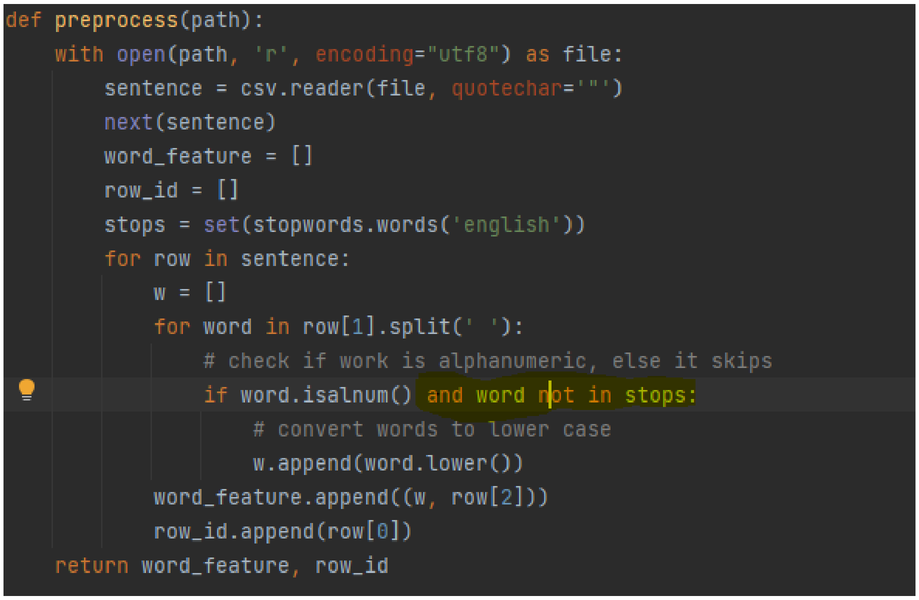
We made the decision to build our classifier with stop words included in the vocabulary as the performance is not improved when stop words are discarded as seen in the confusion mat

Confusion matrix for stop words included in the vocabulary:

System\	Character	Director	Performer	Publisher

ground truth				
Character	80	5	5	2
Director	8	83	5	2
Performer	6	3	91	2
Publisher	9	3	2	94

To reproduce the code without stop word, simply include the following line highlighted in the preprocess function found in the util.py file.



```
def preprocess(path):
    with open(path, 'r', encoding="utf8") as file:
        sentence = csv.reader(file, quotechar='"')
        next(sentence)
        word_feature = []
        row_id = []
        stops = set(stopwords.words('english'))
        for row in sentence:
            w = []
            for word in row[1].split(' '):
                # check if work is alphanumeric, else it skips
                if word.isalnum() and word not in stops:
                    # convert words to lower case
                    w.append(word.lower())
            word_feature.append((w, row[2]))
            row_id.append(row[0])
    return word_feature, row_id
```

Performance Report

Accuracy:

We made use of 3-fold cross-validation for training the dataset. The k-fold cross validation uses different portions of our training data to train the model.

This produced an accuracy average of 0.8820.

Training Accuracy:

N Split = 3	Individual Accuracy from K-fold cross-validation training
Train set 1	0.89655
Train set 2	0.87556
Train set 3	0.87387

We applied the trained model to our test data and produced an accuracy of 0.893.

Test Accuracy

Data	Accuracy
Test	0.893

Accuracy on the test set is far from the average training accuracy which shows the model is overfitting

Confusion Matrix

System\ ground truth	Character	Director	Performer	Publisher
Character	90	6	6	3
Director	6	82	4	1
Performer	4	3	91	2
Publisher	3	3	2	94

Classification Report: micro-and-macro-averaged precision and recall

Features\report	Precision	Recall	Fi-score	Support
Character	0.857	0.874	0.865	103
Director	0.882	0.872	0.877	94
Performer	0.910	0.883	0.897	103
Publisher	0.922	0.940	0.931	100
Accuracy			0.892	400
Macro avg	0.893	0.892	0.892	400
Weighted avg	0.893	0.892	0.892	400

The micro average precision is equal to 0.892 as shown in the table above.

Error Analysis

From the confusion matrix shown above, Characters, performer, and director have the highest misclassification because they contain similarly related words.

Common misclassification is shown below:

Sentence	original	System prediction	Line number	Reason
Odell in particular has mentioned James Hetfield of	performer	Character	417	Word contains

Metallica as his biggest influence in his guitar - work , mostly notably the track " Sad But True " .				relation between a track and an artist which the classifier predicted to be a character.
Earlier that year Witherspoon was chosen to portray June Carter Cash , the second wife of country music singer - songwriter Johnny Cash (Joaquin Phoenix) , in " Walk the Line "	characters	performer	418	Word contains singer, and songwriter, which the classifier predicted to be in the performer relation
His first professional role was as Rum Tum Tugger in the Andrew Lloyd Webber musical " Cats " , in 1990 - 1991 .	characters	director	2590	Word contained musical, and role, which the classifier predicted to be in the director relation

The publisher relation has the least amount of misclassification, this is likely due to the rare words, not being similar to other relations.