# CENG463 Homework-1

Elif Ecem Ümütlü - 2448991

December 17, 2023

**Abstract**

This homework has 2 parts each explained below under Task1 and Task2 titles.

# 1 TASK 1

## 1.1 Preprocessing

The dataset contains title, description pairs for books. For each class (genre), we have train, development, and test datasets separately. In the preprocessing part, below operations applied sequentially.

Throughout this report, "document" will be used to indicate a book (title, description pair for any class).

### 1.1.1 Punctuation removal

For the classification task, we are not interested in the end of the sentences. Our concerns are whether a word occurs in a document or not.

### 1.1.2 Contraction

Informal and formal written language differs from each other. Some characters in some words in the documents are omitted in informal spoken and written language such as he's, won't, etc. In order to get the full form of the words, I used contractions on them. (he's → he is)

### 1.1.3 Noisy word removal

Some words in the documents are not meaningful, some of them are misspelled. Additionally, some misspelled words are included in more than one document. Since I have used frequency of the words for feature selection, those overly repeated misspelled words create a noise; hence, results in the drop of accuracy, f1-score, etc.

I eliminated all words that are not in the dictionary. Dictionary is obtained from the library nltk.corpus.words.

### 1.1.4 Character lowering

Some words occurred in the documents both with first letter capitalized, and lowered. As a result, even though some words occurred more, their frequency is calculated separately.

i.e. Calm calm Calm Calm Calm calm → number(calm) : 2 , number(Calm) : 4

However, it should have been as such: → number([Cc]alm): 6

For this reason, I have lowered down all characters in the words.

### 1.1.5 Lemmatization

In feature selection, I used frequency of the word. Some words occur in the documents in different forms even though they have the same form, word meaning. I applied lemmatization to obtain the base form of the words. i.e. surfing → surf

#### 1.1.6 Adding title twice

Titles summarize the main topic of the books. Hence, the words in them have great importance acquiring the general idea of which genre the book might have belong. For this reason, I added words in the title twice.

#### 1.1.7 Combining in tuple

Those preprocessing techniques mentioned above are applied to both title and description of the documents in each genres. Finally, each (title, description) pair after the text processing is combined with their genre, and added to the megadocument.

Here is an example of the processed file: [('republic republic presented form dialogue socrates three different classic text enquiry notion perfect community ideal individual within conversation raised goodness reality knowledge republic also purpose education role men people remarkable lucidity deft use allegory plato depiction state bound harmony philosopher', 'philosophy'), ... ]

### 1.2 Feature Selection

At the beginning of the homework, I was not aware of the library functions. However, my thinking process has followed such a way:

To classify a document, we use word probabilities. Therefore, word selection has the utmost importance. I decided to choose the word that identifies the class, according to the word meaning and number of occurrence of the word. But doing it by hand was tedious and not efficient. After some research, I have found library functions that apply what I mentioned above. Finally, I used the below functions to utilize feature selection.

#### 1.2.1 TF-IDF

TF-IDF is used for selecting frequently used and unique words in the vocabulary for each class.

TF stands for Term Frequency. It measures the number of times a word occurs in the documents over words in all documents. In other words, it measures the frequency of the word.

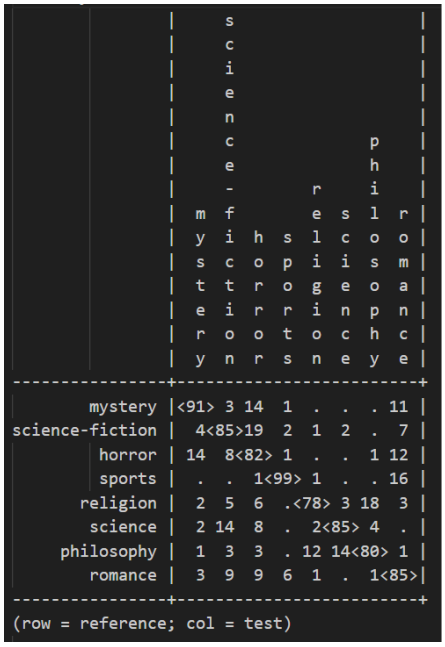
IDF stands for Inverse Document Frequency. Some terms occur in documents of different classes. However, we are interested in unique frequently used words for each genre. These commonly used words in all genres are not much informative. IDF has a lower value for these words.

#### 1.2.2 Chi-square (Chi2)

I used Chi-square to extract the most unique, frequently used K words according to the TF-IDF values. Instead giving a threshold value for p-values of the chi-square result (giving a threshold would be used for null hypothesis rejection, acception), I sorted the p-values and get the most informative 1200 words.

### 1.3 Naive Bayes Classifier

To obtain the train, test, and development sets I applied feature selection algorithm to preprocessed documents. I used nltk function to obtain accuracy, F1 score, recall, and precision. nltk.ConfusionMatrix has a method to print the scores. Accuracy is 71.9 for the test set. Confusion matrix can be seen below (second one belongs to the test set). The diagonals represent the true positives. In my classifier, horror has the lowest F-measure with 0.6308, whereas sports documents are have shown a better performance with 0.8761 F-score. Those values indicates that some selected features are not unique enough to identify a document correctly for horror, or romance. However, documents of mystery, science, and sports are easier to predict with my features.

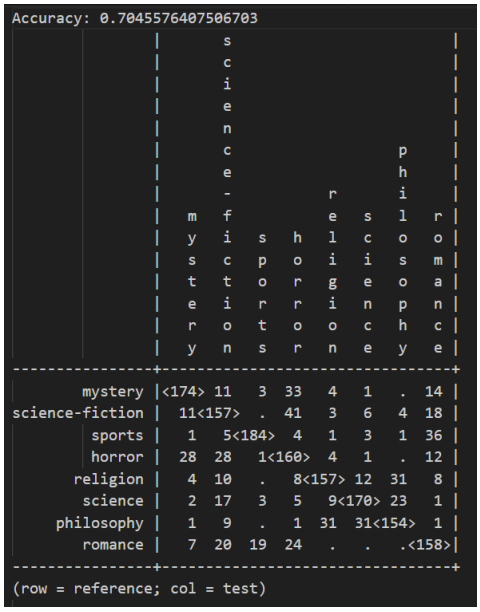


| | mystery | science-fiction | horror | sports | religion | science | philosophy | romance |
|---|---|---|---|---|---|---|---|---|
| mystery | <91> | 3 | 14 | 1 | . | . | . | 11 |
| science-fiction | 4 | <85> | 19 | 2 | 1 | 2 | . | 7 |
| horror | 14 | 8 | <82> | 1 | . | . | 1 | 12 |
| sports | . | . | 1 | <99> | 1 | . | . | 16 |
| religion | 2 | 5 | 6 | .<78> | 3 | 18 | 3 | |
| science | 2 | 14 | 8 | . | 2 | <85> | 4 | . |
| philosophy | 1 | 3 | 3 | . | 12 | 14 | <80> | 1 |
| romance | 3 | 9 | 9 | 6 | 1 | . | 1 | <85> |

(row = reference; col = test)

| Tag | Prec. | Recall | F-measure |
|---|---|---|---|
| horror | 0.5775 | 0.6949 | 0.6308 |
| mystery | 0.7778 | 0.7583 | 0.7679 |
| philosophy | 0.7692 | 0.7018 | 0.7339 |
| religion | 0.8211 | 0.6783 | 0.7429 |
| romance | 0.6296 | 0.7456 | 0.6827 |
| science | 0.8173 | 0.7391 | 0.7763 |
| science-fiction | 0.6693 | 0.7083 | 0.6883 |
| sports | 0.9083 | 0.8462 | 0.8761 |

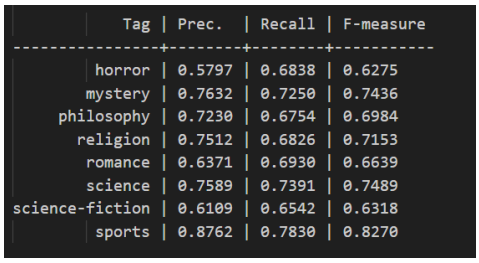
Figure 1: Confusion Matrix of Test Set - NBC



Accuracy: 0.7045576407506703

| | mystery | science-fiction | sports | horror | religion | science | philosophy | romance |
|---|---|---|---|---|---|---|---|---|
| mystery | <174> | 11 | 3 | 33 | 4 | 1 | . | 14 |
| science-fiction | 11 | <157> | . | 41 | 3 | 6 | 4 | 18 |
| sports | 1 | 5 | <184> | 4 | 1 | 3 | 1 | 36 |
| horror | 28 | 28 | 1 | <160> | 4 | 1 | . | 12 |
| religion | 4 | 10 | . | 8 | <157> | 12 | 31 | 8 |
| science | 2 | 17 | 3 | 5 | 9 | <170> | 23 | 1 |
| philosophy | 1 | 9 | . | 1 | 31 | 31 | <154> | 1 |
| romance | 7 | 20 | 19 | 24 | . | . | . | <158> |

(row = reference; col = test)



| Tag | Prec. | Recall | F-measure |
|---|---|---|---|
| horror | 0.5797 | 0.6838 | 0.6275 |
| mystery | 0.7632 | 0.7250 | 0.7436 |
| philosophy | 0.7230 | 0.6754 | 0.6984 |
| religion | 0.7512 | 0.6826 | 0.7153 |
| romance | 0.6371 | 0.6930 | 0.6639 |
| science | 0.7589 | 0.7391 | 0.7489 |
| science-fiction | 0.6109 | 0.6542 | 0.6318 |
| sports | 0.8762 | 0.7830 | 0.8270 |

Figure 2: Confusion Matrix of Test Set - SVC

### 1.3.1 Data Analysis

After I have trained my model with training set, I used development set to get the common most used words between two missclassifed classes (like horror and mystery). In my next evaluation, I subtracted those common word from my features and trained the model accordingly. After that, I got a higher accuracy in both test and development set.

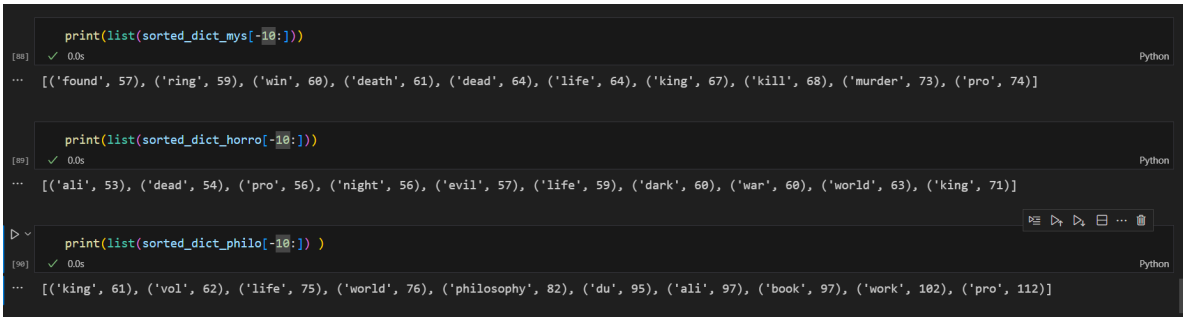



Figure 3: Feature evaluation using development set

By inspecting the confusion matrix, I observed that mystery, romance, and science-fiction book may not be well-classified. Additionally, mystery books can be misclassified as horror, or romance, and visa versa. Moreover, philosophy books may be misclassified as religion, or science, and visa versa.

### 1.4 Support Vector Machine Classifier

The same feature selection is applied to the base data. This means, the same train, development, and test sets are used to evaluate the models and predictions. The accuracy was lower than that of naive bayes classifier, with 0.7024 accuracy.

Similar to the naive bayes classifier, horror has the lowest F-measure value, and sports has the highest F-measure value.

As opposed to what I think, SVC underperformed the Naive Bayes classifier. I was expecting SVC to achieve higher scores since naive bayes only uses the word frequency. Maybe, feature selection process should be different for SVC.

## 2 TASK 2

### 2.1 Question 1

Second column is used to match words with their part-of-speech tags. Third one is used to indicate more informative features of POS tag. For example, when second column is VERB, third column can be VBN, VB, VBZ, VBP, VBG, VBD, each of which gives different information about the verb.

VBN: Verb Past Participle

VB: Verb in base form

VBZ: Verb, Third Person Singular Present

...

We are using the third column for this task.

### 2.2 Question 2

| Set number | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Set1 | 0.9144 | 0.9144 | 0.9144 | 0.9144 |
| Set2 | 0.9267 | 0.9267 | 0.9267 | 0.9267 |
| Set3 | 0.9270 | 0.9270 | 0.9270 | 0.9270 |

#### 2.2.1 Set2

In addition to the set1 features, I used previous word, and suffices of the current word. Since suffix length can be 1, 2, or 3, I added words' last 1, 2, and 3 characters.

#### 2.2.2 Set3

In addition to the set2 features, I used the knowledge of whether previous and next words contain any digit or not.

### 2.3 Question 3

In our case, I used previous word as a feature since some part-of-speech tags are formed based on them. For example, before a date (NUM - CD), it is likely that the previous word is in (IN). Therefore, word dependency is important for feature selection. Logistic regression relates words with their environments. As opposed to LR, CRF does not refrain from knowledge checking of word dependency. Hence, CRF should be preferred for this task.