Classification Assignment

Group 2

Supervised by: DR. Arya Rahgozar

Choosing data:

First we did some web scrapping to get 5 books which are (Alice’s adventures in wonderland: fantasy book), (Dracula: fantasy book), (Emma: Romantic/ Fantasy), (the war of the worlds: Fantasy), and (Uncle Tom’s cabin or life among the lowly: Fantasy /Novel of manners)

We wanted to have a variety of topics within the same genre so we chose all of them to be fantasy but we have added some extra genres like Romantic/Fantasy in Emma’s book in order to make our model confused so it will be well trained

Preprocessing and Data Cleansing:

Then we started with preprocessing as we removed the special characters, stop words and the unusual spaces as cleaning for the data.

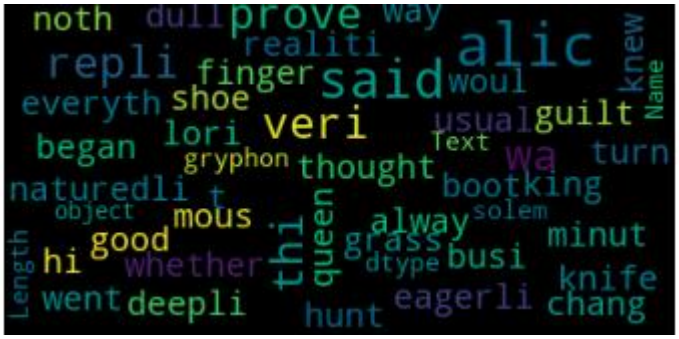
Now we have our data prepared so we grouped our data by the name of the book and we have encoded the labels using LabelEncoder.

To be able to deal with our data we transformed it BOW and TF-IDF and n-gram.

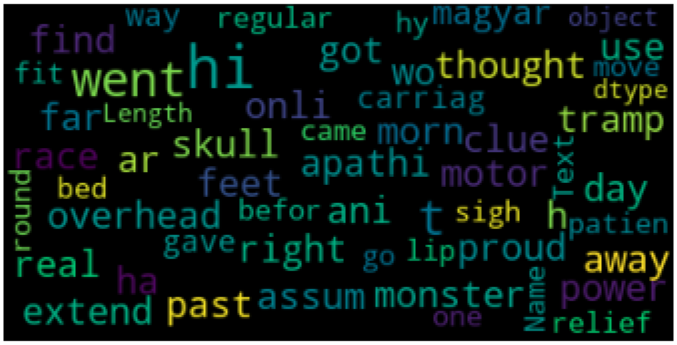
To reduce the number steps in our code and to have a better shape for the data, we used the pipeline technique.

To visualize our data we did **EDA** “Exploring data analysis”, by visualizing the most repeated 100 word for each class.

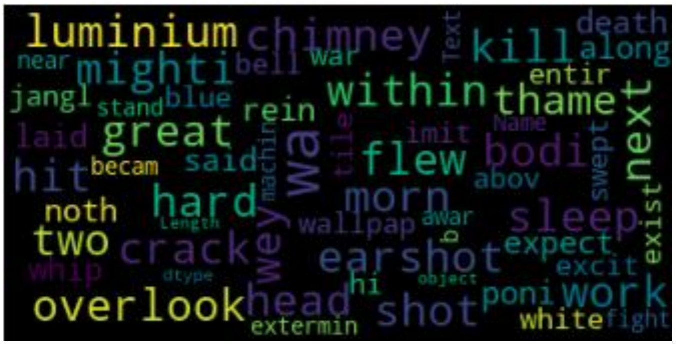
1) Top words for Alice’s adventures in wonderland



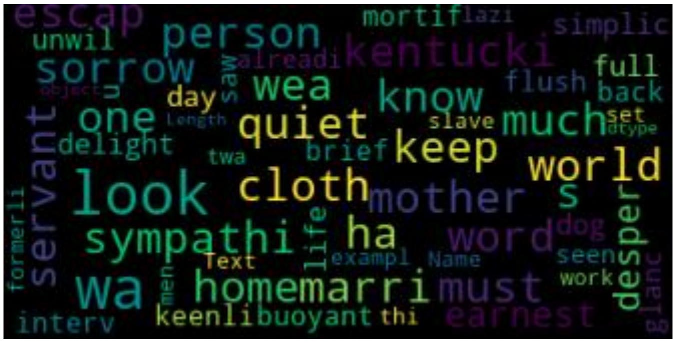
2) Top words for Dracula



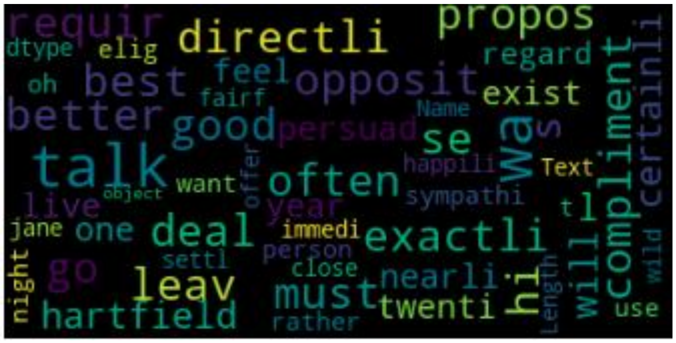
3) Top words for War of the Worlds



4) Top words for uncle Tom’s



5) Top words for Emma



Feature Engineering:

We have used pipeline technique to make our code easier to be manipulated and then we have tested them with multiple combination of feature engineering techniques like “BOW”, “TF-IDF”, “N-grams” on different models so we can determine the best feature engineering combination to contribute in raising our accuracy the most.

Using SVM, Decision Tree, k-Nearest Neighbor, and Naïve Bayes:

We have used four models (SVM, Decision Tree, k-Nearest Neighbor, and Naïve Bayes).

And as we said to be able to use those models we used the “pipeline” technique.

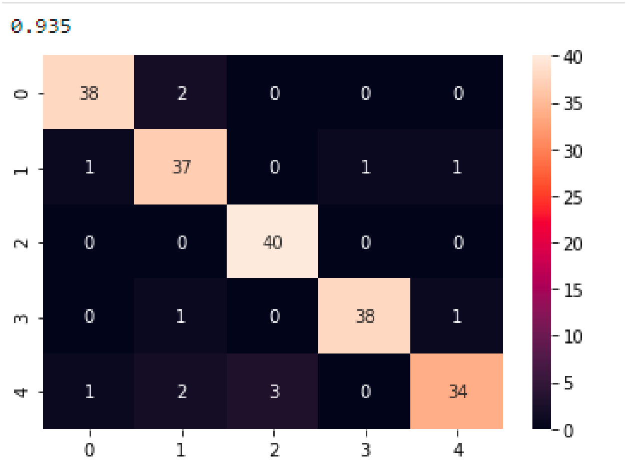
Then we started to train our four models with different way of transformation each time.

Then we made an evaluation applying the cross validation to our four models to obtain our champion model.

We had two champion models which are the Naïve Bayes and the SVM, because each time we run our code the generated data are always random and the accuracy of both models is almost the same.

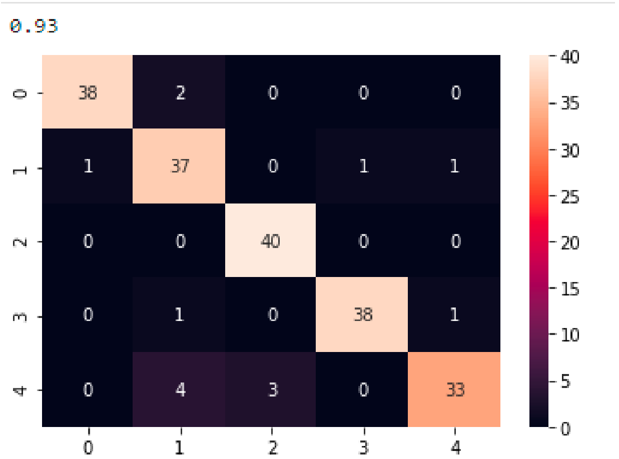
Naïve Bayes

Naïve Bayes with BOW



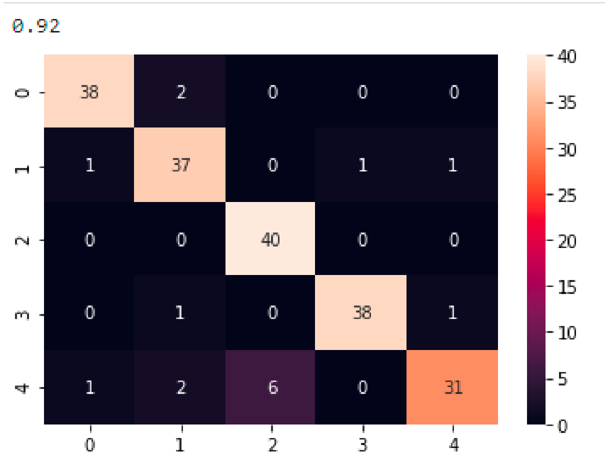
```
0.935
```

Naïve Bayes with BOW and N-gram



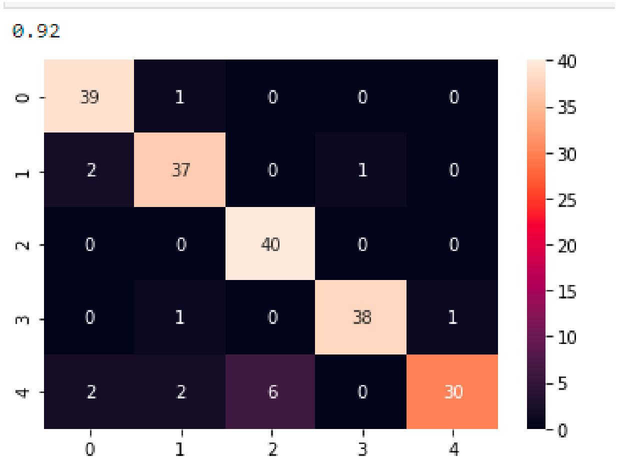
```
0.93
```

Naïve Bayes with TF-IDF



```
0.92
```

Naïve Bayes with TF-IDF and N-gram

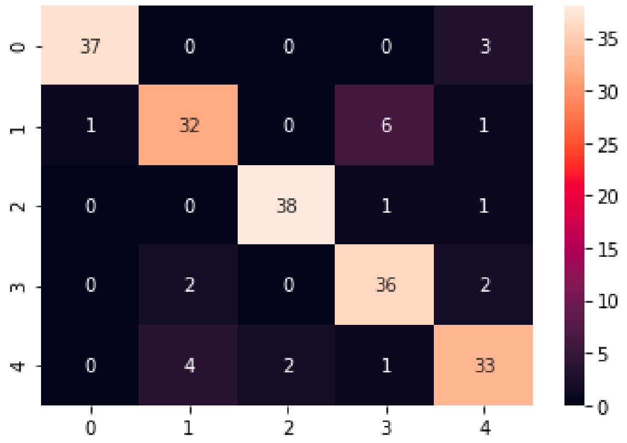


```
0.92
```

SVM

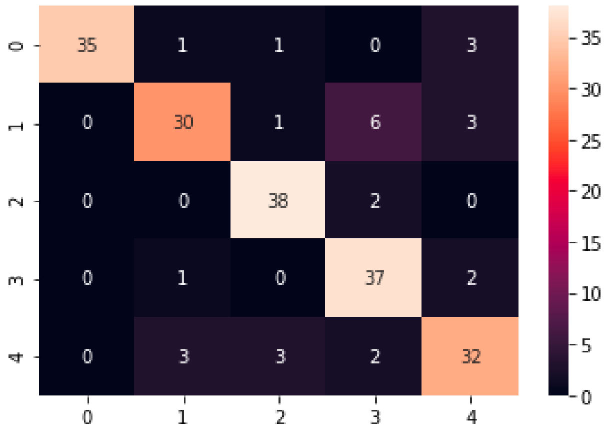
SVM with BOW

0.88



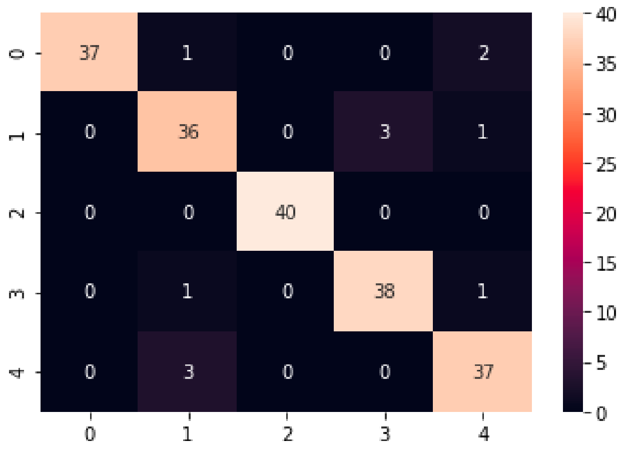
SVM with BOW and N-grams

0.86



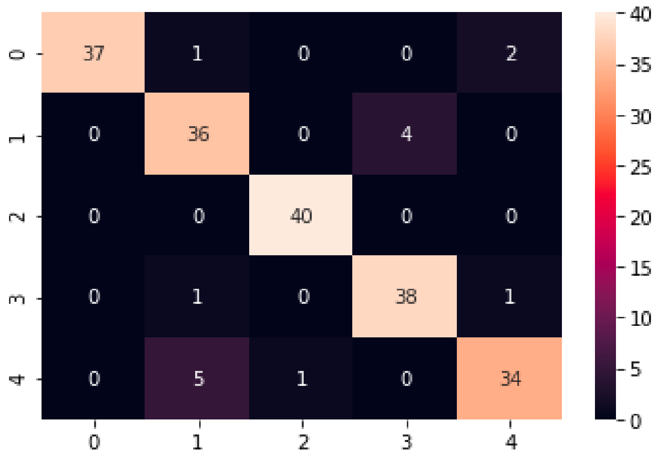
SVM with TF-IDF

0.94



SVM with TF-IDF and N-grams

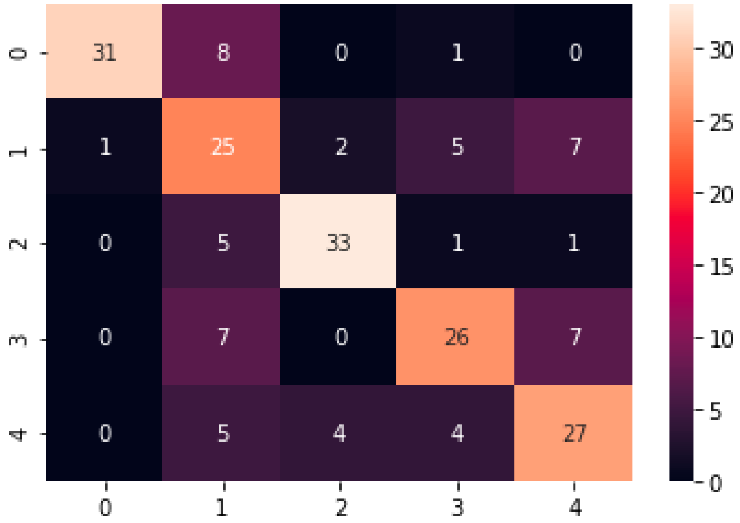
0.925



Decision tree

Decision tree model with BOW

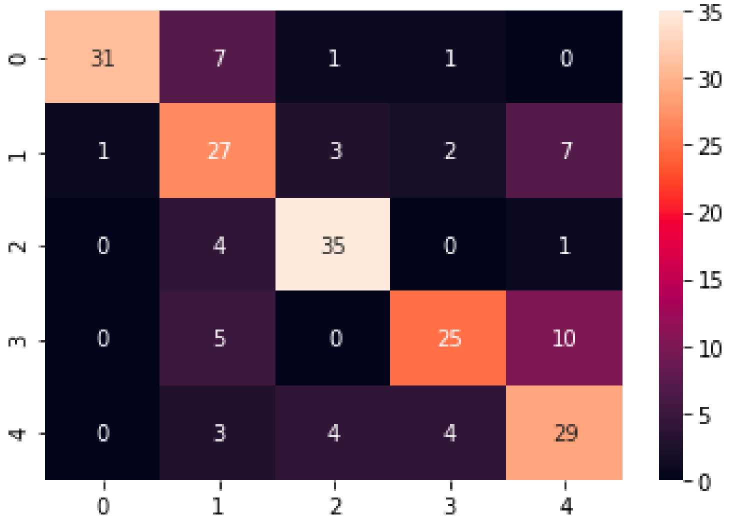
```
0.71
```



	0	1	2	3	4
0	31	8	0	1	0
1	1	25	2	5	7
2	0	5	33	1	1
3	0	7	0	26	7
4	0	5	4	4	27

Decision tree model with BOW and N-gram

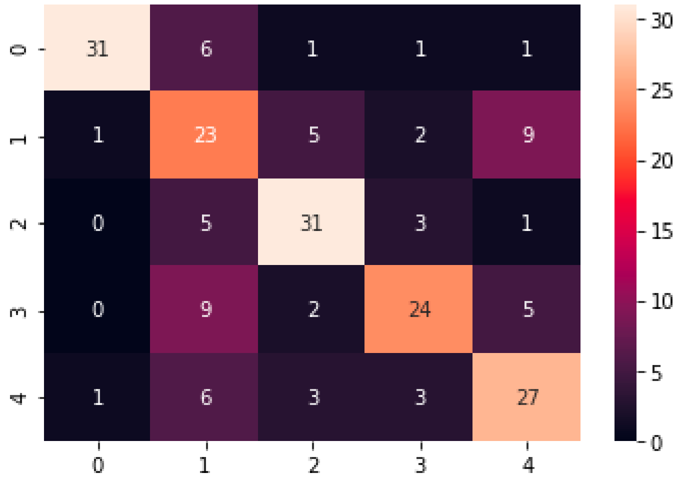
```
0.735
```



	0	1	2	3	4
0	31	7	1	1	0
1	1	27	3	2	7
2	0	4	35	0	1
3	0	5	0	25	10
4	0	3	4	4	29

Decision tree with TF-IDF

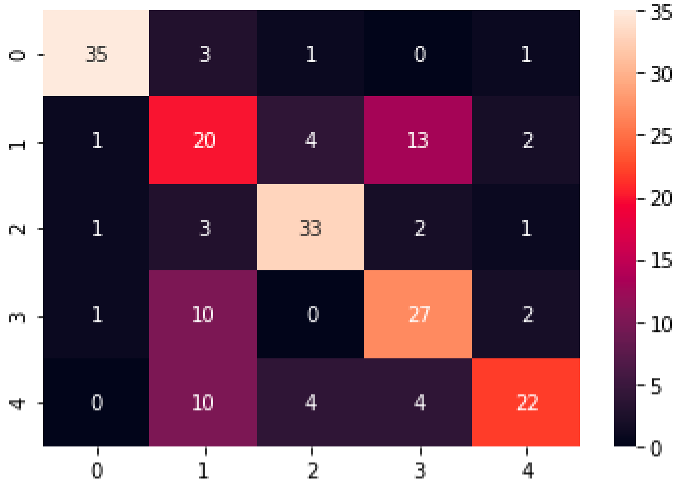
```
0.68
```



	0	1	2	3	4
0	31	6	1	1	1
1	1	23	5	2	9
2	0	5	31	3	1
3	0	9	2	24	5
4	1	6	3	3	27

Decision tree with TF-IDF N-gram

```
0.685
```

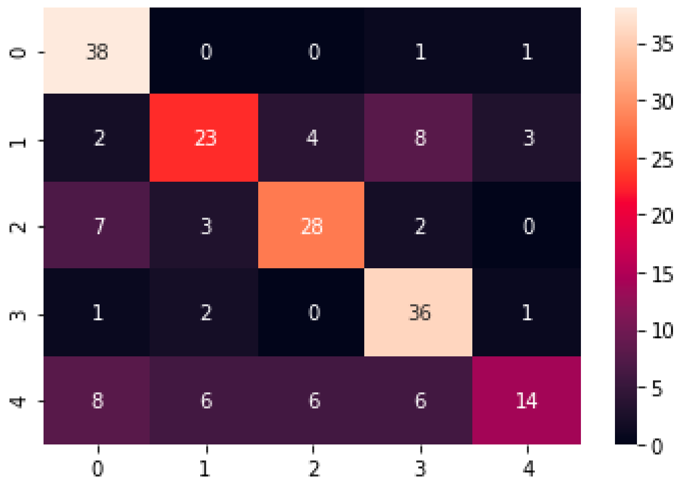


	0	1	2	3	4
0	35	3	1	0	1
1	1	20	4	13	2
2	1	3	33	2	1
3	1	10	0	27	2
4	0	10	4	4	22

KNN

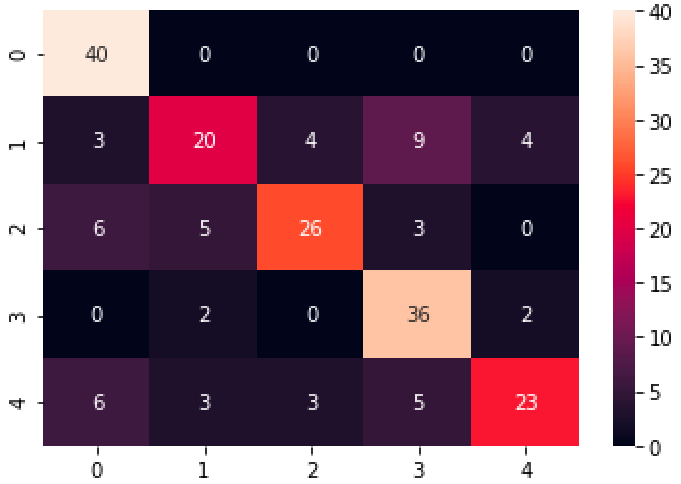
KNN model with BOW

```
0.695
```



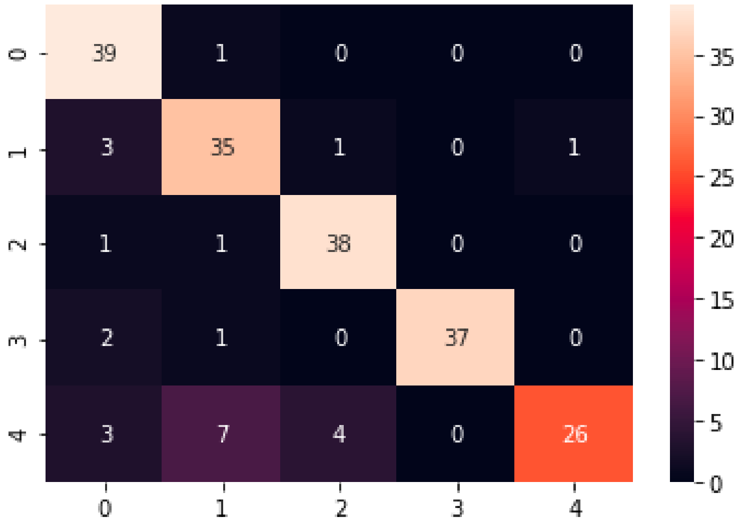
KNN model with BOW and N-grams

```
0.725
```



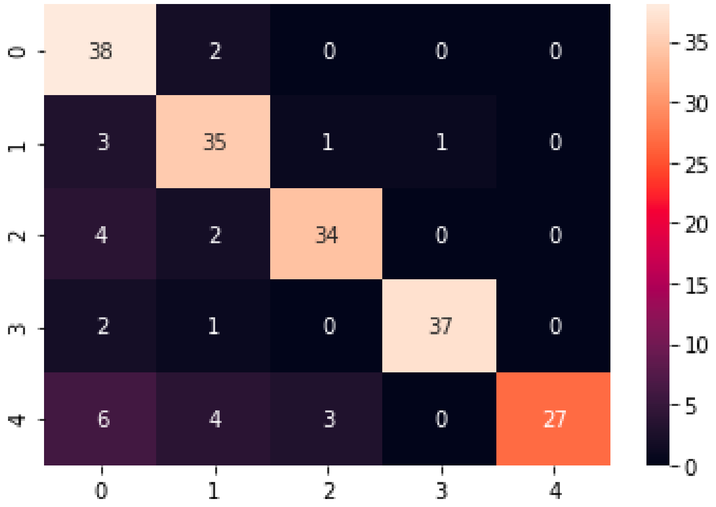
KNN with TF-IDF only

```
0.875
```



KNN with TF-IDF N-grams

```
0.855
```

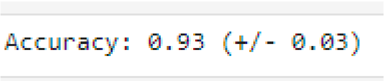


Perform Evaluations and Analysis of Bias and Variability:

We have used Cross-Validation

1) Naïve Bayes

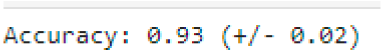
   a) With using: CountVectorizer, MultinomialNB



```
Accuracy: 0.93 (+/- 0.03)
```

We have got 93% Accuracy and 0.03 Standard deviation.

   b) With using: CountVectorizer, N-gram_, and MultinomialNB

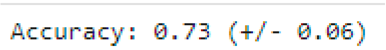


```
Accuracy: 0.93 (+/- 0.02)
```

We have got 93% Accuracy and 0.02 Standard deviation.

2) Decision Tree

With using: CountVectorizer, and DecisionTreeClassifier

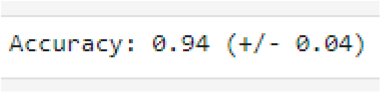


```
Accuracy: 0.73 (+/- 0.06)
```

We have got 73% Accuracy and 0.06 Standard deviation.

3) SVC

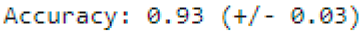
   a) With using: CountVectorizer,TfidfTransformer,and SVC.



```
Accuracy: 0.94 (+/- 0.04)
```

We have got Accuracy 94% and 0.04 Standard deviation.

   b) With using: CountVectorizer, N-gram, TfidfTransforme,and SVC

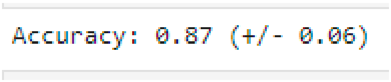


```
Accuracy: 0.93 (+/- 0.03)
```

We have got Accuracy 93% and 0.03 Standard deviation.

4) KNN

With using: CountVectorizer(ngram_range ,TfidfTransformer,KNeighborsClassifier)

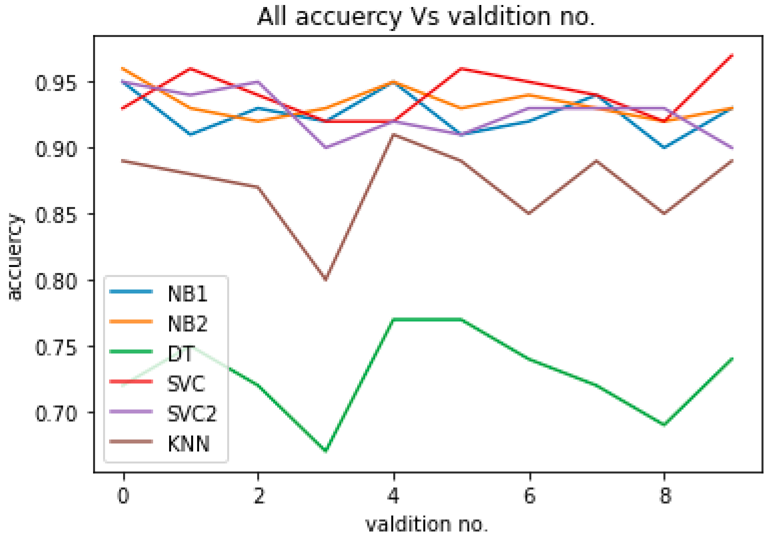


```
Accuracy: 0.87 (+/- 0.06)
```

We have got Accuracy 87% and 0.06 Standard deviation.

So we have decided that the **Champion model** is Naïve Bayes with BOW and N-gram as their feature engineering procedures

Although SVC have a relatively high mean and variance but it's over all accuracy was 89%, that's why we chose the Naïve Bayes a both validation and the test accuracy were relatively high

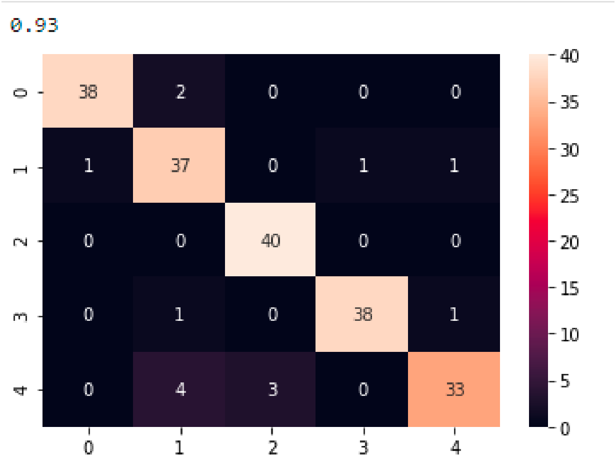


Perform Error Analysis

When we observed the confusion matrix and the classification part we found that the class Emma had a bad precision accuracy because there was a lot of false positive in it, and Uncle Tom’s book also had a bad recall.

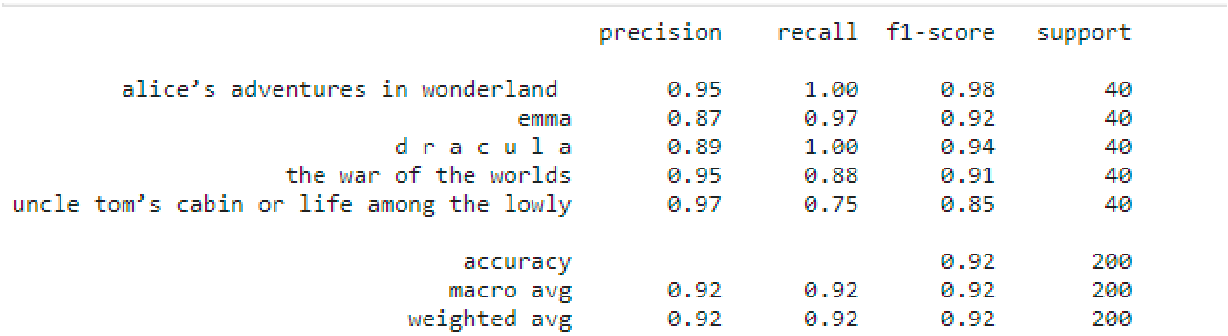
There were a lot of similarity between Emma and the other books which we have discovered from the precision that we have received from our champion model, so the model labeled some of “Emma’s” documents as another book because Emma’s precision were low.

Uncle Tom’s recall accuracy which we have received with our champion model was very low with 75% which mean only 75% of Uncle Tom’s documents were actually true as we predicted.



```
0.93
```

	0	1	2	3	4
0	38	2	0	0	0
1	1	37	0	1	1
2	0	0	40	0	0
3	0	1	0	38	1
4	0	4	3	0	33



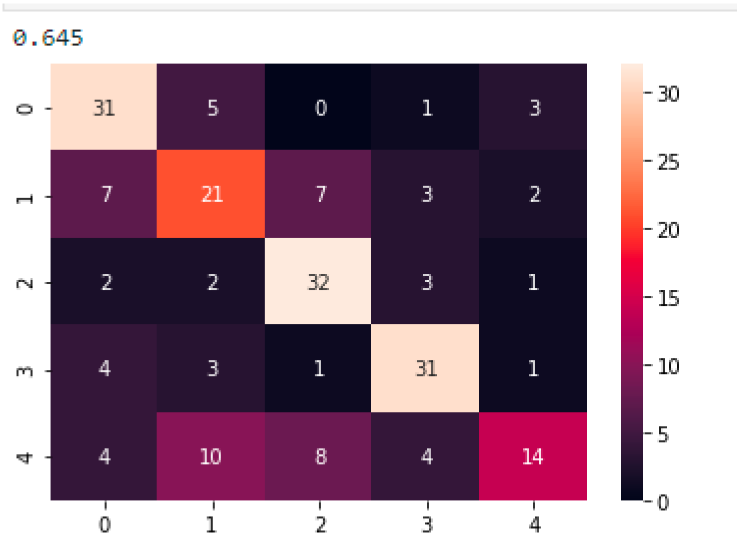
```
                                        precision    recall  f1-score   support

      alice’s adventures in wonderland       0.95      1.00      0.98        40
                                  emma       0.87      0.97      0.92        40
                         d r a c u l a       0.89      1.00      0.94        40
                  the war of the worlds       0.95      0.88      0.91        40
uncle tom’s cabin or life among the lowly       0.97      0.75      0.85        40

                              accuracy                           0.92       200
                             macro avg       0.92      0.92      0.92       200
                          weighted avg       0.92      0.92      0.92       200
```

Bring the accuracy down

We have limited the words in each document to be only 20 words instead of 100 words and we got 65%



```
0.645
```

	0	1	2	3	4
0	31	5	0	1	3
1	7	21	7	3	2
2	2	2	32	3	1
3	4	3	1	31	1
4	4	10	8	4	14