

Applications of Data Science

EBC 5125

Assignment 1

Submitted By

Vamsi Kiran Tamanala (300095158)

Venkata Sai Krishna Reddy (300110183)

Prahasith Kondra (300143154)

Bharat Battula (300110122)

**1 Data Preparation:**

Step1: Importing the required libraries and downloading the Gutenberg books.

|  |  |
| --- | --- |
| Input  Output |  |

Step2: Importing the seven textbooks from the Project Gutenberg Books.

|  |  |
| --- | --- |
| Reading the 7 books as the text files |  |

Step3: Creating an array for the Authors

|  |  |
| --- | --- |
| Authors |  |

Step 4: Removing the stop words:

Stop words are removed using the RegX, Lower Case, Full Stop.

|  |  |
| --- | --- |
| 1 |  |

|  |  |
| --- | --- |
| 2 |  |

Step 5: Labelling the documents with the corresponding authors:

|  |  |
| --- | --- |
| Labelling |  |

Step 6: Adding all the documents into a single Corpus

|  |  |
| --- | --- |
| Corpus |  |

**2 Data Processing**

Converting the data into pandas

|  |  |
| --- | --- |
| Data  Frame |  |

Data Visualization:

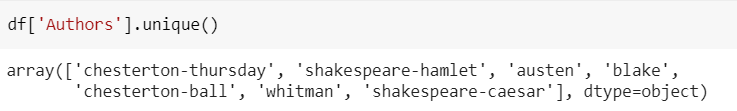
|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

Counting the number of sentences that each other has in the entire corpus:

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |



**Feature Engineering:**

|  |  |
| --- | --- |
| X is the test data  y is the target data |  |

**3 Transformation of the Data**

1.Transformation of the data using Bag of Words (BOW):

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output |  |

Splitting the data into train and test data using BOW:

|  |
| --- |
|  |

2.Transformation of the data using TF-IDF:

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output |  |

Splitting the data into train and test data using TF-IDF:

|  |
| --- |
|  |

Classification of the data using the algorithms

1.Multinomial Naïves Bayes Model

2.Decision Tree

3.Support Vector Machine

4. K Nearest Neighbors

**1.Multinomial Naives Bayes Model:**

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

Testing the model:

Testing the model with BOW:

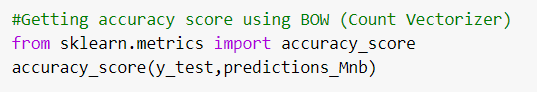
|  |  |
| --- | --- |
| Input |  |

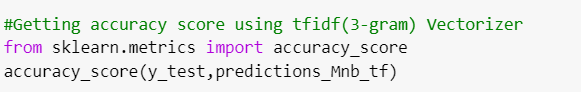
|  |  |
| --- | --- |
| Output and Validation  Accuracy: 0.77 |  |

Testing the model with TF- IDF:

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation  Accuracy: 0.72 |  |





**2.Decision Tree:**

|  |
| --- |
|  |
|  |
|  |

Testing the model:

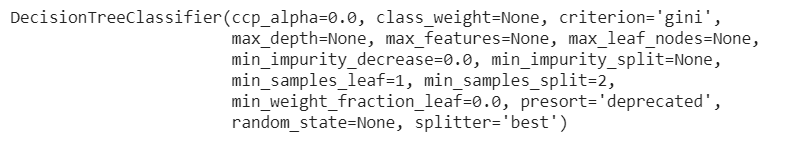
Testing the model with BOW:

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation  Accuracy: 0.61 |  |

Testing the model with TF-IDF;

|  |  |
| --- | --- |
| Input |  |



|  |  |
| --- | --- |
| Output and Validation  Accuracy:  0.6121517393357468 |  |

**3. Support Vector Machine:**

|  |
| --- |
|  |
|  |

Testing and Validating the model with BOW:

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation  Accuracy: 0.685502912010074 |  |

Testing the model with TF-IDF

|  |
| --- |
|  |
|  |

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation:    Accuracy: 0.68691956555957 |  |

**4. K Nearest Neighbors:**

|  |
| --- |
|  |
|  |

Testing the model with BOW:

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation    Accuracy: 0.43459782779789075 |  |

Testing the model with TF-IDF:

|  |
| --- |
|  |
|  |

|  |  |
| --- | --- |
| Input |  |

|  |  |
| --- | --- |
| Output and Validation:    Accuracy: 0.17047064379033527 |  |

**7 Results:**

|  |  |  |
| --- | --- | --- |
| Algorithm | BOW | TF-IDF |
| Multinomial NB | 0.77 | 0.72 |
| Decision trees | 0.61 | 0.6121517393357468 |
| SVM | 0.685502912010074 | 0.68691956555957 |
| KNN | 0.43459782779789075 | 0.17047064379033527 |