CS777 – Week 5 Homework Submission Template

**!!!! PLEASE RENAME THIS DOCUMENT WITH YOUR NAME AND LASTNAME !!!!**

**Task 1 – Vectorize the Data**

* Vectorize the text documents using the available functions in the MLlib

|  |
| --- |
| countVectorizer = CountVectorizer(inputCol=remover.getOutputCol(),  outputCol="features\_c", vocabSize=5000,minDF=10, maxDF=5000) |

* Print the first ten words of the vocabulary

|  |
| --- |
| *['tribunal', 'applicant', '', 'ltd', 'respondent', 'appellant', 'j', 'pennsylvania', 'pty', 'squadron']* |

* Print the total time needed to vectorize the data

|  |
| --- |
| *0.018196582794189453* |

**Task 2 – Using the Logistic Regression Model**

* Print the F1-measure and **confusion matrix**

|  |
| --- |
|  |

* Print the total time needed to train the model

|  |
| --- |
| *train the model: 272.8136818408966 secs*  *Evaluate the model: 83.24093580245972 secs*  *Test the model: 0.2566392421722412 secs*  *Total Time: 356.31125688552856 secs* |

* Calculate the performance metrics

|  |
| --- |
| *Precision: 0.999199999063587*  *Recall: 0.9991988891262551* |

**Task 3 – Using the SVM Model**

* Print the F1-measure and **confusion matrix**

|  |
| --- |
|  |

* Print the total time needed to train the model

|  |
| --- |
| *train the model: 279.35752034187317 secs*  *Evaluate the model: 67.25431251525879 secs*  *Test the model: 0.2057032585144043 secs*  *Total Time: 346.81753611564636 secs* |

* Calculate the performance metrics

|  |
| --- |
| *Precision: 0.9995741325656741*  *Recall: 0.999572740867336* |

**Task 4 – Feature Selection**

* Describe your feature selection technique and discuss why you chose it

|  |
| --- |
| *I used Chi-squared, it works well for text data and run faster than PCA* |

* Discuss if your dimension reduction approach is applicable to very large data sets

|  |
| --- |
| *I think so, since I got the results. But for even larger data, it is highly parallelizable, suitable for distributed computing* |

* Repeat tasks 2 and 3 with reduced features and report the F1 measure, confusion matrix, and computation time

|  |
| --- |
| *Logistic Regression:* |

|  |
| --- |
| *SVM:* |

* Comment on the differences in the results

|  |
| --- |
| *The results did not show much differences, very slightly improvement, but the running time increased a lot.* |

**Spark History Output:**

**Task 1:**

|  |
| --- |
|  |

**Task 2:**

|  |
| --- |
| *I run all tasks together* |

**Task 3:**

|  |
| --- |
| *Your screen capture should go here.* |