CS777 – Week 4 Homework

**Task 1 – Data Preparation**

* Print the frequency position of the words “applicant”, “and”, “attack”, “protein”, and “car”

|  |
| --- |
| *Index for 'applicant' is 448*  *Index for 'and' is 2*  *Index for 'attack' is 512*  *Index for 'protein' is 3167*  *Index for 'car' is 648*  *Index for 'in' is 3* |

**Task 2 – Learning the Model**

* Print the five words with the largest (positive or negative) regression coefficients

|  |
| --- |
| *Five words most strongly related to an Australian court case: ['that', 'court', 'mr', 'applicant', 'tribunal']* |

* Print how your cost is changing for each iteration

|  |
| --- |
| *Iteration 0 Loss: 117062.8547359383*  *Iteration 1 Loss: 156315.7096237602*  *Iteration 2 Loss: 149137.25016011982*  *Iteration 3 Loss: 142837.4870688273*  *Iteration 4 Loss: 137357.02625729408*  *Iteration 5 Loss: 132736.48119185847*  *Iteration 6 Loss: 129200.59037180113*  *Iteration 7 Loss: 127097.79627924143*  *Iteration 8 Loss: 126351.76368901663*  *Iteration 9 Loss: 126195.23270672705*  *Iteration 10 Loss: 126130.06698415722*  *Iteration 11 Loss: 126077.99071535989*  *Iteration 12 Loss: 126033.16646221436*  *Iteration 13 Loss: 125994.38879478443*  *Iteration 14 Loss: 125960.8154425953*  *Iteration 15 Loss: 125931.73360367463*  *Iteration 16 Loss: 125906.53169816073*  *Iteration 17 Loss: 125884.68387817149*  *Iteration 18 Loss: 125865.73741887609*  *Iteration 19 Loss: 125849.30211507124*  *Iteration 20 Loss: 125835.04131858476*  *Iteration 21 Loss: 125822.66433821316*  *Iteration 22 Loss: 125811.91997694646*  *Iteration 23 Loss: 125802.59102252698*  *Iteration 24 Loss: 125794.48954025241*  *Iteration 25 Loss: 125787.45284326667*  *Iteration 26 Loss: 125781.34003681308*  *Iteration 27 Loss: 125776.0290501404*  *Iteration 28 Loss: 125771.41408379687*  *Iteration 29 Loss: 125767.40341156647*  *Iteration 30 Loss: 125763.91748579957*  *Iteration 31 Loss: 125760.88730276308*  *Iteration 32 Loss: 125758.25299118842*  *Iteration 33 Loss: 125755.96259267913*  *Iteration 34 Loss: 125753.97100724166*  *Iteration 35 Loss: 125752.23908108303*  *Iteration 36 Loss: 125750.73281709738*  *Iteration 37 Loss: 125749.4226912482*  *Iteration 38 Loss: 125748.2830604165*  *Iteration 39 Loss: 125747.2916493036*  *Iteration 40 Loss: 125746.42910569902*  *Iteration 41 Loss: 125745.67861490039*  *Iteration 42 Loss: 125745.02556533503*  *Iteration 43 Loss: 125744.45725852039*  *Iteration 44 Loss: 125743.96265743168*  *Iteration 45 Loss: 125743.53216815216*  *Iteration 46 Loss: 125743.15745036969*  *Iteration 47 Loss: 125742.8312528841*  *Iteration 48 Loss: 125742.54727080236*  *Iteration 49 Loss: 125742.30002154544* |

**Task 3 – Evaluation of the Learned Model**

* Print the computed F1 score obtained by your classifier and the **confusion matrix.**

|  |
| --- |
| *Performance Metrics: Logisitic Regression with Gradient Descent*  *Precision: 0*  *Recall: 0.0*  *F1: 0*  *Confusion Matrix:*  *TP: 0 , FP: 0*  *TN: 165373 , FN: 3513*  *Accuracy: 0.9791989862984498* |

* Look at the test samples for three false positives or false negatives and describe why you think your model was fooled

|  |
| --- |
| *It's notable that the model did not produce any false positives (FP: 0), which means that it did not incorrectly classify any non-Australian court cases as Australian ones. However, it's worth noting that the model also did not identify any true positives (TP: 0), indicating that it failed to correctly identify any Australian court cases at all. This is further evidenced by the F1 score of 0, which suggests that the model's precision and recall were both ineffective in this context.*  *Given that there are no false positives to analyze, we can't directly address why the model might have been fooled in certain instances. However, we can speculate on why the model might struggle in general based on the provided metrics and common challenges in text classification.*  *One potential issue could be class imbalance, as evidenced by a significantly higher number of true negatives (TN: 165373) compared to false negatives (FN: 3513). If the model was trained on a dataset with a similar imbalance, it might be biased towards predicting the majority class (non-Australian court cases).*  *Another potential issue could be the features used to train the model. If the model was trained on features that are not strongly indicative of the target class, it might struggle to make accurate predictions. For instance, if the model was trained on common legal terms that appear in both Australian and non-Australian court cases, it might not be able to effectively distinguish between the two.*  *Lastly, the model might be too simplistic to capture the complexities of the data. Logistic Regression, while effective for linearly separable data, might not be able to capture more complex patterns in the data that could be leveraged to more accurately predict the target class.*  *In summary, while we cannot analyze specific false positives in this instance, potential reasons for model misclassification in general could include class imbalance, ineffective feature selection, and model simplicity. Future work might involve addressing these issues, for instance, by resampling the data, engineering more effective features, or exploring more complex models.* |

**Spark History Output:**

**Task 1:**

|  |
| --- |
| *spark_history* |

**Task 2:**

|  |
| --- |
| *spark_history* |

**Task 3:**

|  |
| --- |
| *spark_history* |