CS777 – Week 4 Homework Submission Template

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

**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

|  |
| --- |
| *['an', 'but', 'been', 'time', 'about']* |

* Print how your cost is changing for each iteration

|  |
| --- |
| *Iteration No. 0 Cost= 17.669133886603245*  *Iteration No. 1 Cost= 17.056276789875575*  *Iteration No. 2 Cost= 16.852440430030303*  *Iteration No. 3 Cost= 16.685631249263423*  *Iteration No. 4 Cost= 16.540389984346064*  *Iteration No. 5 Cost= 16.410745907344417*  *Iteration No. 6 Cost= 16.293321425510165*  *Iteration No. 7 Cost= 16.185844953795712*  *Iteration No. 8 Cost= 16.086645354443984*  *Iteration No. 9 Cost= 15.994429303798924*  *Iteration No. 10 Cost= 15.908164640889822*  *Iteration No. 11 Cost= 15.82701113101027*  *Iteration No. 12 Cost= 15.750275440936571*  *Iteration No. 13 Cost= 15.677379835562203*  *Iteration No. 14 Cost= 15.607839330523674*  *Iteration No. 15 Cost= 15.541244415288563*  *Iteration No. 16 Cost= 15.477247642353609*  *Iteration No. 17 Cost= 15.415553010821798*  *Iteration No. 18 Cost= 15.355907435153*  *Iteration No. 19 Cost= 15.29809380991446* |

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

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

|  |
| --- |
| *--Contingency matrix--*  *TP: 0 FP: 0*  *FN: 377 TN: 18347*  *----------------------*  *Accuracy = 0.9798654133732109*  *F1 = 0.0*  *Precision = 0*  *Recall = 0.0* |

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

|  |
| --- |
| *The dataset is extremely imbalanced, I tried both weighted and not weighted function with different parameters but still can not get good results. In that way, the model could only find the TN but not TP. I may say model cannot find the words that can represent legal cases.* |

**Spark History Output:**

**Task 1:**

|  |
| --- |
|  |

**Task 2:**

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

**Task 3:**

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