CS777 – Week 4 Homework Submission Template

**Task 1 – Data Preparation**

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

|  |
| --- |
| *Your output should go here.* |

* Include the relevant code excerpt that you used to create an RDD for the dictionary that includes the 10,000 most frequent words

|  |
| --- |
| *The code excerpt should go here.* |

**Task 2 – Learning the Model**

* Print the five words with the largest regression coefficients

|  |
| --- |
| *Your output should go here.* |

* Print how your cost is changing for each iteration

|  |
| --- |
| *Your output should go here.* |

* Include the relevant code excerpt that you used for converting each document in the training set to a vector with 10,000 entries

|  |
| --- |
| *The code excerpt should go here.* |

* Include the relevant code excerpt that you used for learning a logistic regression model for this task

|  |
| --- |
| *The code excerpt should go here.* |

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

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

|  |
| --- |
| *Your output should go here.* |

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

|  |
| --- |
| *Your explanation should go here.* |

* Include the relevant code excerpt that you used for computing the F1 score and confusion matrix

|  |
| --- |
| *The code excerpt should go here.* |

**Spark History Output:**

**Task 1:**

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

**Task 2:**

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

**Task 3:**

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