Kartik Garg

Sentient Science: Data Engineer

UB Alumni

ADIC Assignment Documentation

Question 1:

1. Part A.

The input file was course\_data.csv which contains marks of students in a course with their gender.

According to the analysis using the boxplot we can see that overall Female ratio of marks scored in the course is higher than the male ratio.

Part B.

The input file box\_office\_data.csv contains a list of: Movies and their Ids with COSTS (Million\_USD) and REVENUE (Million\_USD).

The analysis shows a bar plot with profit margin as a sort argument and the plot displays the the movie with highest profit margin was “It” and the one with the lowest profit margin was “King Arthur: Legend of the Sword”.

Part C.

For part C the input file was ncdc\_data.csv. It contains a list of all the days in the year 1990 with their precipitation and temperature.

The analysis contains a “Heat Wave vs Time” graph in 30 days margin.

Question 2:

1. This file demonstrates the use of pdfminer library in python. The input are 3 pdf file beaches.pdf, Gettysburg.pdf and pearl\_harbor.pdf.

Pdfminer library reads the pdf into text which hels in further analysis of the text.

Further analysis of the text from “beaches.pdf” shows:

('the', 63)

('and', 33)

('of', 33)

('to', 19)

('we', 15)

('shall', 13)

('in', 12)

('a', 9)

('that', 9)

('on', 8)

As the top 10 words used in the pdf

Analysis of \*gettysburg.pdf\* shows:

('the', 13)

('that', 10)

('to', 8)

('of', 8)

('a', 7)

('we', 6)

('not', 5)

('for', 5)

('here', 5)

('can', 5)

As the top 10 words used in the pdf.

Analysis of \*pearl\_harbor.pdf\*:

('the', 41)

('of', 24)

('and', 16)

('to', 12)

('that', 12)

('Japanese', 10)

('our', 9)

('will', 8)

('in', 8)

('United', 6)

Words and their iterations common in all three pdf's

('and', 1)

('all', 1)

('is', 1)

('it', 1)

('are', 1)

('have', 1)

('in', 1)

('our', 1)

('for', 1)

('their', 1)

Question 3:

1. Tweepy package is used to read all the twitter feeds and do an analysis on them. In my analysis we will try to find

Tweets mentioning Cancer and various products which are mentioned with the word

“Cancer”

----------------Tweets mentioning Cancer and various products------------------

coffee : 3

beer : 5

meat : 0

sugar : 41

cigarettes : 2

nuts : 0

sunscreen : 0

chips : 0

Total\_Mentions : 51

The word “Sugar” was mentioned the highest number of times with the word “Cancer” following “Coffee” and “cigarettes”. This can give us an insight as to what people feel about when they think about the word cancer.

Question 4:

1. This model uses pytorch library to predict origin or words/names.

The mnist\_model saves the training set, which further helps to predict the names.

This is what we get as a result:

name = Popescu, origin = Czech

name = Fernandez, origin = Portuguese

name = Velenzuela, origin = German

name = Lovecraft, origin = French

name = Chambers, origin = English

name = Davies, origin = Portuguese

name = Paltrowski, origin = Polish

name = Sargiannis, origin = Greek

name = Ovechkin, origin = Russian

name = Fapp, origin = German