Name – Navarurh Kumar  
NET ID – NXK180010

**FINAL TERM PROJECT FOR DATABASE FOUNDATIONS**

**Objective** – Utilize data and gain information based on business questions.

**Technology Used**

Operating System – Ubuntu 16.04 LTS

Database – MySQL

Languages – Python(for data cleaning), R(for cleaning and loading to MySQL)

**Understanding and Cleaning Data**

The first step in the process was to get the raw data files and analyze them for their formatting and their discrepancies to make sure that the data is in good shape before loading it into a database. The following steps outline the process followed and notes (in red) explain the discrepancies found -

1. The data in BUAN6320-DataSet3.txt had a lot of special characters that were causing issues when importing the data into R for cleaning. The removal of these characters was carried out using a python script to clean out characters by forcing UTF-8 encoding for the file. The script used was -

x = open("datasets/BUAN6320-DataSet3.txt", "r")

opt = x.read()

x.close()

opt = opt.decode('utf-8','ignore').encode("utf-8")

print opt

o = open("datasets/BUAN6320-DataSet3\_pytmp.txt", "w")

o.write(opt)

o.close