Northeastern University

Course:	CS6020
Assignment:	Module 4 - Data Import - D
Total Points:	100
Date Due:	Posted on Blackboard

Learning Objectives

In this assignment, you will learn how to:

- read and parse text files

Tasks

1. (100 Points) Load and then parse the IMDB movie listing from the file movies.list.gz. Note that the file is compressed so you need to figure out how to uncompress it. Inspect the file and determine how to best load it -- this is not an XML file and requires custom string parsing. Place the data into a data frame suitable for further analysis.

   Only read in the movies. You can ignore TV shows and the "porn" movies -- they have a special marking character. Make any other assumptions you need, but comment your assumptions.

   Perhaps build a smaller subset of the file that's easier for testing and loads faster. This is a common technique when building data loaders.