# Northeastern University

| | |
|---|---|
| **Course:** | DA5020 |
| **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. (30 points) Load 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 in R. Inspect the file and determine how to best load it -- this is not an XML file and requires custom string parsing.
2. (20 points) Parse the data. You should identify all the fields and their meanings within the file. Place the data into a data frame suitable for further analysis.
3. (10 points) Comment your code where you identify the movie rows that are part of your result set.
4. (20 points) Your result set should only contain movie title and movie release year. Your result set should **NOT** include rows for TV shows. You can identify the movies within the data file ( look for a special marking field or some other indication). Make any other assumptions you need, but comment your assumptions. For a cleaner result set, look for duplicate titles and remove the duplicates.
5. (20 points) For correct syntax, coding style and readable code format.

If the data size is overwhelming, you can build a smaller subset of the file that's easier for testing, loads faster and is representative of the file ( a random sample of rows). This is a common technique when building data loaders. This technique is for debugging purposes only. Your submitted assignment should be run against the complete data set.