**CSC 571-485B / SENG 480A: Summer 2017**

**Assignment 3**

**Due: June 15, 2017, 11:55 pm**

Implement the following exercises using MapReduce and Spark.

Consider the movielens dataset provided with the partial code (see folder “movielens”).

See: <https://grouplens.org/datasets/movielens> for a description of the dataset.

1. (1 pt) Find the average rating for each user and movie.

Use MapReduce in Hadoop.

Use the “ratings.csv” file. This file is copied to “input\_movielens” subdirectory.

Use the same MR program to compute both in one go.

Use a combiner (additionally to mapper and reducer).

2. (3 pt) Find the average rating for each genre.

Use MapReduce in Hadoop.

Use the “movies.csv” file to find the genre of each movie.

This file is assumed to fit in the memory of a machine; it is cached locally.

A movie can belong to multiple genres.

A rating of a movie should be used for all the genres the movie belongs in.

Use a combiner (additionally to mapper and reducer).

3. (1 pt) Repeat 1 using Spark.

4. (5 pt) Repeat 2 using Spark.

The partial code and Hadoop and Spark can be downloaded from: <http://webhome.cs.uvic.ca/~thomo/assign3_hadoop_spark.zip>

(see the README file)

**Submit all your source code and your outputs (but not Hadoop and Spark; they are too big for Connex).**