For this in-lab assignment, you will be analyzing the movie ratings data that we collected in earlier in the semester. A starter project will be provided that produces the following interface, populates a data structure that holds the student names and ratings, and performs objective 1) in the menu below.

Minimally for this project, you need to complete objective 2) in addition to one of the other objectives (and you can make up your own objective if you like in 7) and 8)). For 6) you should look to a student’s movie friend for advice.

Note that if a sufficient number of groups finish this project on Friday, we will have time for (short) presentations on Monday, May 8th.

Mathematical Background:
Assume that each student $S_i$ has a distribution $R_i$ of movie ratings with mean $\mu_i$ and standard deviation $\sigma_i$ (we write $R_i \sim (\mu_i, \sigma_i)$ to indicate this). Our goal is to use this information (along with a few assumptions) to compute the Pearson Correlation Coefficient $\rho_{ij} \in [-1,1]$ for each pair of students $S_i$ and $S_j$ (note this computation is symmetric, i.e. $\rho_{ij} = \rho_{ji}$). A given student’s movie friend is the student for whom the Pearson Correlation Coefficient is the largest, and their movie enemy is the student from whom this coefficient is the smallest (note that you must exclude a given student in their own evaluation of a movie friend, since $\rho_{ii} = 1$ for all $i$.)

Note that since we only take into account movies that both users have seen in the computation of this coefficient, being a movie friend is generally not going to be symmetric (that is, your best movie friend may be a particular student, and their best movie friend does not have to necessarily be you). This is because not everyone rated all of the movies.
The **Pearson Correlation Coefficient** of the distributions $R_i \sim (\mu_i, \sigma_i)$ and $R_j \sim (\mu_j, \sigma_j)$ of two student’s $S_i$ and $S_j$ rankings is formally defined as follows:

$$
\rho_{ij} = \frac{E[(R_i - \mu_i)(R_j - \mu_j)]}{\sigma_i \sigma_j}
$$

where the operator $E$ denotes the expected value of the expression. While this expression is difficult to calculate in practice, we make the following simplifying assumption:

1. A particular student’s $S_i$ average ranking $\mu_i$ does not change appreciably if we restrict this average to ONLY the movies that they have in common with student $S_i$. Thus, as we compare student $S_i$ with other students, we will assume that this value is a constant $\mu_i$.

To approximate the numerator of this expression, we let $r_{ik}$ and $r_{jk}$ be the rankings of student $S_i$ and student $S_j$ for a movie $k$ that they have both rated. In particular, if $M$ denotes the number of movies that these two students have both rated, then the approximation for the numerator of $\rho_{ij}$ is obtained by taking the average of the $M$ products $(r_{ik} - \mu_i)(r_{jk} - \mu_j)$

$$
E[(R_i - \mu_i)(R_j - \mu_j)] \approx \frac{1}{M} \sum_k (r_{ik} - \mu_i)(r_{jk} - \mu_j)
$$

Similarly, we approximate the standard deviations in the denominator of $\rho_{ij}$ by averaging the squares of the deviations of the rankings from the respective means, but we again only include movies for which both students have rated:

$$
\sigma_i \approx \sqrt{\frac{\sum_k (r_{ik} - \mu_i)^2}{M}} \text{ and } \sigma_j \approx \sqrt{\frac{\sum_k (r_{jk} - \mu_j)^2}{M}}
$$

Putting this all together, with some algebraic simplification we can calculate the Person Correlation Coefficient between students $S_i$ and $S_j$ as

$$
\rho_{ij} = \frac{\sum_k (r_{ik} - \mu_i)(r_{jk} - \mu_j)}{\sqrt{[\sum_k (r_{ik} - \mu_i)^2] \ast [\sum_k (r_{jk} - \mu_j)^2]}}
$$

where in both the numerator and denominator, the summation is ONLY over the movies $k$ that both students $S_i$ and $S_j$ have rated.

**NOTE:** It will be helpful to note that if a student did not rank a particular movie, then their ranking for that movie has been set to 0. Thus, if the product of the rankings of a particular movie for a pair of students is zero, then this means that at least one of the students did not rank the movie and thus it should not be included in the computation of $\rho_{ij}$ above.
Example:

Student 0: Ranks movies (1, 2, 3, 4, 5)
Student 1: Ranks movies (5, 4, 3, 2, 1)
Student 2: Ranks movies (1, 0, 0, 0, 5)

Notice that the students have overall averages of $\mu_0 = 3, \mu_1 = 3, \text{ and } \mu_2 = 3$, respectively.

Consider students 0 and 2. Because these students both only ranked the 1st and 5th movies, these are the only movies that we consider in the calculation of the correlation coefficient. In the numerator, we sum over all products of the form $(r_{0k} - \mu_0)(r_{1k} - \mu_1)$ where $r_{0k}$, for example, indicates student 0’s ranking for movie k.

Similarly, in the denominator for each student we compute the sum of the squares of the deviations of the individual movie rankings from their respective means, multiply these two sums together, and take the square root. This gives the following result:

$$\rho_{02} = \frac{(1-3)(1-3)+(5-3)(5-3)}{\sqrt{[(1-3)^2+(5-3)^2][[(1-3)^2+(5-3)^2]]}} = \frac{8}{8} = 1$$

Note that this is not surprising since students 0 and 2 exactly agreed on the movies that they both ranked.

Consider now students 0 and 1. Notice that while the two students agreed on the 3rd movie, the trend of the ratings are diametrically opposed. Since both of these students rated all movies, our expression will be more complicated:

$$\rho_{01} = \frac{(1-3)(5-3)+(2-3)(4-3)+(3-3)(3-3)+(4-3)(2-3)+(5-3)(1-3)}{\sqrt{[(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2][(5-3)^2+(4-3)^2+(3-3)^2+(2-3)^2+(1-3)^2]]}} = -1$$

Similarly, for students 1 and 2 we have:

$$\rho_{12} = \frac{(5-3)(1-3)+(1-3)(5-3)}{\sqrt{[(5-3)^2+(1-3)^2][(1-3)^2+(5-3)^2]]}} = \frac{-8}{8} = -1$$

Thus:

Student 0: Movie friend is Student 2, Movie enemy is Student 1
Student 1: Either has two movie friends or two movie enemies or neither! **
Student 2: Movie friend is Student 0, Movie enemy is Student 1

**NOTE: For our dataset, all students have a definitive movie enemy and a definitive movie friend, so cases like student 1 do not occur.

A reminder: The correlations coefficients are symmetric, i.e. $\rho_{ij} = \rho_{ji}$ but the idea of a movie friend may not be symmetric!
The Visual Studio Solution contains a number of files:

- **Student.h/cpp** - a concrete class that represents a Student and that holds their ID and first and last name in addition to their overall movie rating average (needed to compute correlations) and their individual movie rankings (in a vector) and correlation coefficients with other students (in a vector).
- **main.cpp** – a driver that contains helper functions to process and display the data.
- **movies.txt** – a text file that consists of the number of movies in the file numMovies on the first line, followed by that many movieID MovieName pairs.
- **students.txt** – a text file whose first line gives number of students numStudents in the file, followed by that many studentID, student first name, student last name triples.
- **ratings.txt** – a text file that contains only movie ratings, one line per student, with individual movie ratings separated by a space. The first row corresponds to the student with ID 0, the second to student with ID 1, etc. The first column corresponds to the ratings for the movie with ID 0, etc. **If the user did not rate a movie, a score of 0 has been entered.**

You may modify these files in whatever way you want, and include additional files as you see fit (for example, a Movie class may be of use, depending on which questions you answer). The underlying data structure is a vector of pointers to Students called **studentVector**.

**What to submit:**
I will pull your project either from Grace (name your project folder cs380s17hw7-<teamName>) or GitHub (in that order). Also, submit a color, double-sided hard copy of all of your code.