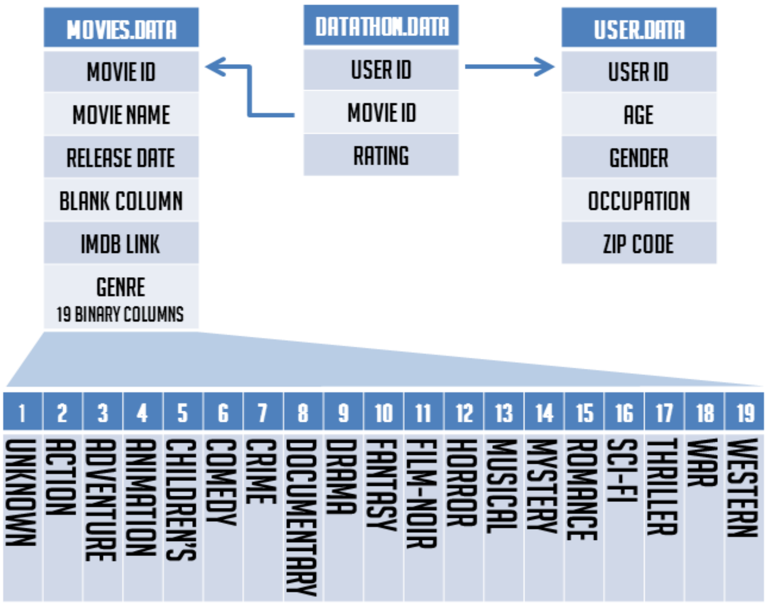
INTRODUCING THE AIB DATATHON

The Challenge

The AIB Datathon will consist of creating a movie recommendation model based on an actual history of user's movie ratings.

The Dataset

The dataset we are using comes from approximately 100,000 movie ratings from almost 1000 users. There is also reference data available for each of the movies and demographic data for each of the users. The dataset contains multiple pipe (|) or tab (\t) delimited records. In addition to the movie ratings, the dataset has information on the genres of the individual movies as well as the sex, age and occupation information for all users. A breakdown of the relationships in the dataset content is shown in the image below.



MOVIES.DATA
MOVIE ID
MOVIE NAME
RELEASE DATE
BLANK COLUMN
IMDB LINK
GENRE
19 BINARY COLUMNS

DATATHON.DATA
USER ID
MOVIE ID
RATING

USER.DATA
USER ID
AGE
GENDER
OCCUPATION
ZIP CODE

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
UNKNOWN	ACTION	ADVENTURE	ANIMATION	CHILDREN'S	COMEDY	CRIME	DOCUMENTARY	DRAMA	FANTASY	FILM-NOIR	HORROR	MUSICAL	MYSTERY	ROMANCE	SCI-FI	THRILLER	WAR	WESTERN

The movie ratings are broken down as follows:

Datathon.data:

Approximately 99,000 records containing the movie id for films watched by each the users (user id) and the ratings they gave each from 1-5(rating).

Datathon_test.data:

This file contains a list of 10 films for 100 test users randomly chosen from the original dataset. For these records we have omitted the ratings they gave.

You should have these files on the USB provided, as well as 'movies.data' and 'user.data'

Building a Recommendation Model

The modelling challenge will involve creating a movie recommendation model based on the dataset. What underpins a good recommender is the ability to predict the degree to which a user will like or dislike a particular film they haven't seen. The task here today is to take a set of 100 users and analyse the rating history we have available for them to predict what scores they would give other films. We have set aside 10 films for each of the 100 users for which we want a prediction made for the rating.

There are many different approaches which could be employed to do this using the given dataset.

What constitutes a Model?

Generating random numbers between 1 and 5 would be a model (kind of, not a good one but a model). We are going to assume that our user ratings aren't random so the following are a few simple models that will perform better than random.

Averages

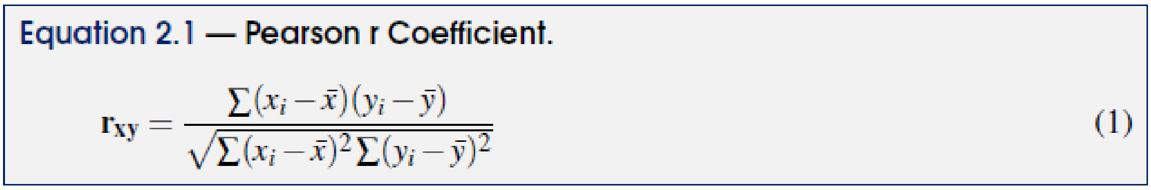
Simply assume that every film in the test set will have a rating close to the average rating for all films in the dataset. This will (nearly always) be an improvement on random...but we can do better.

Grouped Averages

We could assume that every film in the test set list will have a rating close to the average score for that film obtained from the ratings of other users in the training set. We could get a more granularly grouped average by looking at the average over added demographics such as sex & occupation. This is an improvement on the previous model.

Weighted Similarity Model

This is a good bit trickier but there is a simple example case explained below. The next bit has some maths in it but hopefully it's fairly clear what's going on. The method weights user ratings based on the similarity between the User and all the other users in the dataset. It does so by examining the Pearson r correlation coefficient between the User and each other user, calculated based on user submitted ratings of the same movies. This is a number between -1 and 1. If the two users give the exact same ratings to the movies they have watched in common then they will have an r value of 1. Similar users will have r values close to 1. The formula for calculating r is given below

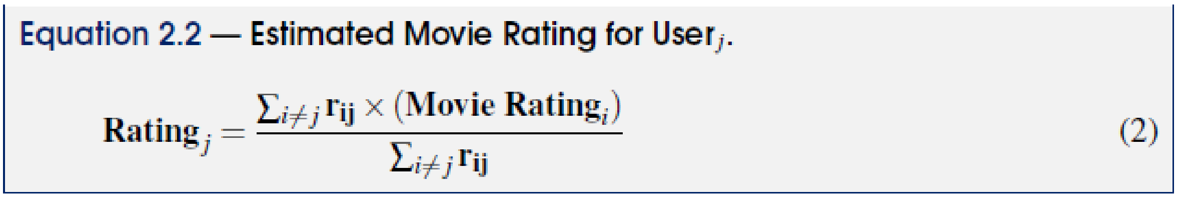


Equation 2.1 — Pearson r Coefficient.

$$\mathbf{r}_{\mathbf{xy}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad (1)$$

Here $\bar{x}$ and $\bar{y}$ denote the average values of x and y respectively. In a movie ratings model, x_i and y_i would be the individual movie ratings given by the two different users. This value must be calculated for each user in the dataset. This will result in a series of similarity measures for each user, describing how similar that user is to every other user in the dataset.

Using this information it is possible to estimate the movie rating the User would give the movies he hasn't seen using Eqn. 2.2. This will return an estimated rating between 0 and 5 for the User.



Equation 2.2 — Estimated Movie Rating for User$_j$.

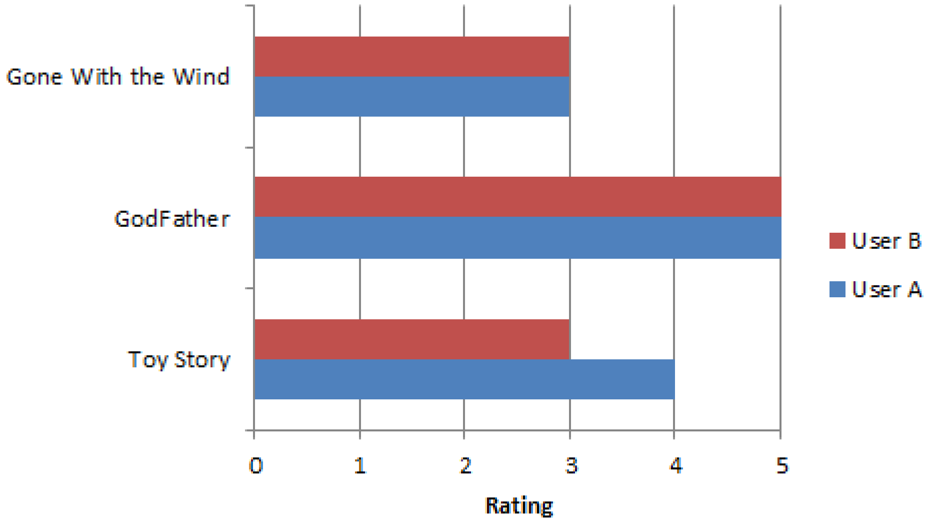
$$\mathbf{Rating}_j = \frac{\sum_{i \neq j} \mathbf{r_{ij}} \times (\mathbf{Movie\ Rating}_i)}{\sum_{i \neq j} \mathbf{r_{ij}}} \qquad (2)$$

where (**Movie Rating**$_i$) is the rating user$_i$ gave for the movie we are investigating. By repeating this calculation for each movie in the database it is possible to construct a recommended list of movies for the User.

You are encouraged to improve this model. Information such as movie genre or the user's sex/age/occupation isn't taken into account here. These could be incorporated into to produce more accurate recommendations.

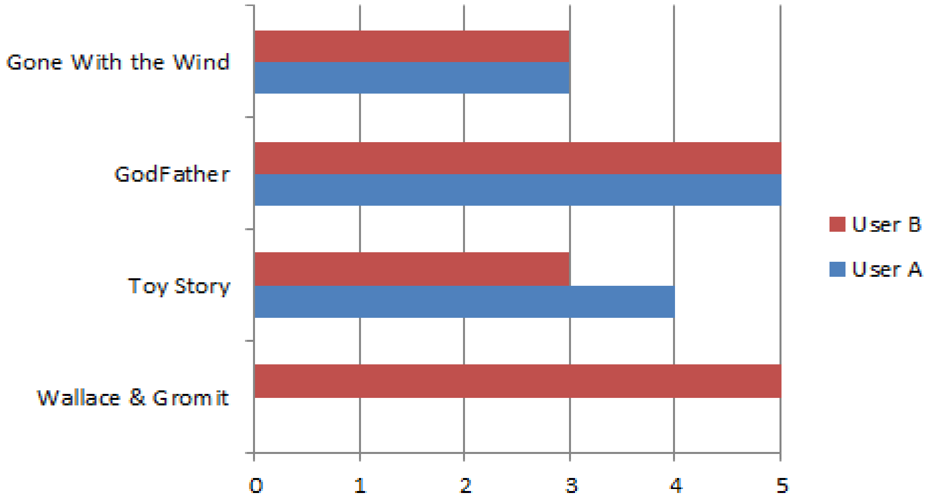
Weighted Similarity (Simple Example)

Let’s say test user A has seen 3 films in common with user B, namely, "Gone With the Wind" , "The Godfather" and "Toy Story", and has rated them 3, 5 and 4. Our other user B has rated these same films as 3, 5 and 3.



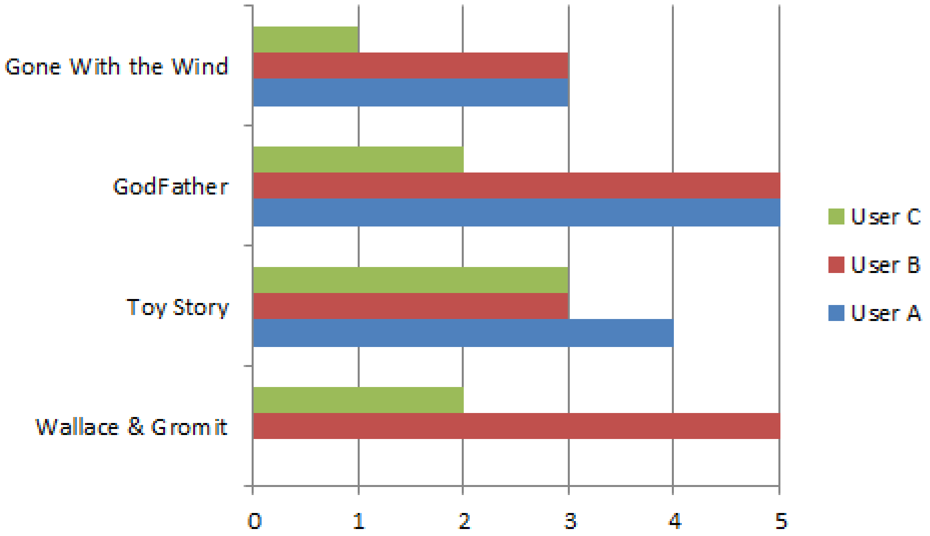
We can see from the chart that user A and user B are quite similar. How similar? It depends on what measure you use but correlation calculated using the equation mention earlier works fairly well. Using this measure they score a similarity of 0.87(remember 1 is the highest possible, -1 the lowest)

Making a prediction



So now if we know that user B has also watched Wallace & Gromit and we are looking to calculate what prediction user A might give that film, then we have a weight that tells how much we value user B’s opinion of that film. User B thinks the film is a 5 and we value his opinion as 0.87 so the predicted score based on this one comparison would be:

The above is an example of a comparison with one user but of course our dataset has many more users. So how do we manage to factor all those guys in? To demonstrate I'll extend our model to include another user, the imaginatively named: 'User C'. He has also watched the same 4 movies as user B so therefore also has the same 3 films in common with User A.



This time though his ratings are less similar for those 3 films as User A's. The correlation value here turns out to be 0.5. So now when we are considering User C's opinion we attribute it less weight.

Now we have a total weight score for Wallace & Gromit of:

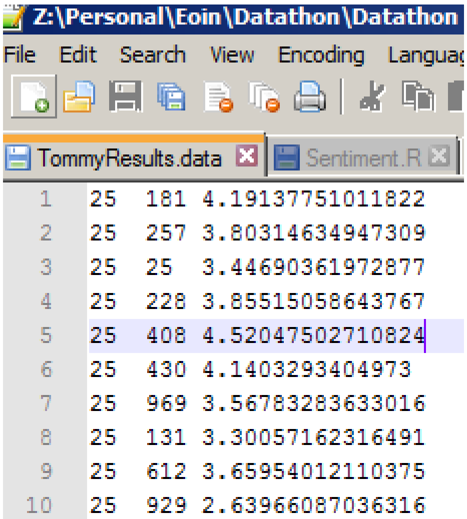
5.35? But the max rating is 5?.... There's one more step we must perform. We now divide by the sum of all the correlation values we calculated along the way. So all in all it works out to be:

Keeping Tabs on your progress & Final Submissions

Modelling:

So now you've built a model, how good is it? When we have made our predictions for the ratings of the 1000 films in the test set, we now want to check how accurate it is versus reality. To that, submit a tab delimited file containing 3 columns: The 'user id', 'movie id' and the 'predicted rating'. For the accuracy scorer to work the file must be **tab delimited (\t)** and also have **NO header**.

Example of predictions for user 25 in the test set. (NOTE: Ratings are always recorded as integers but we do not require that you convert your predictions to integers, we'll leave that up to you)

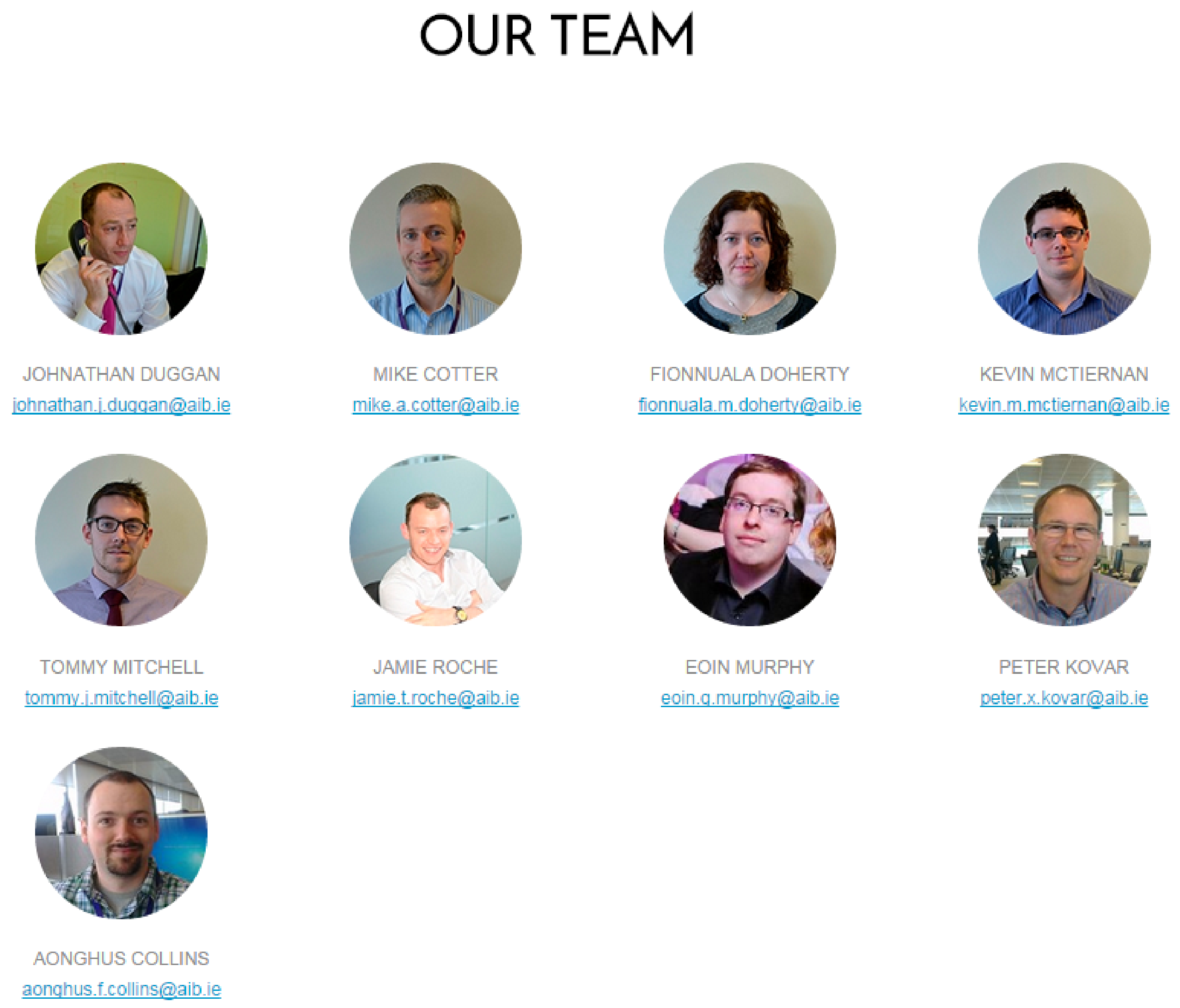


```
25	181	4.19137751011822
25	257	3.80314634947309
25	25	3.44690361972877
25	228	3.85515058643767
25	408	4.52047502710824
25	430	4.1403293404973
25	969	3.56783283633016
25	131	3.30057162316491
25	612	3.65954012110375
25	929	2.63966087036316
```

Once you have a results file and you are interested in finding your score simply upload the results file to your Dropbox Results folder and notify one of the Datathon team that you have done so. Your results will be scored and given a value between 0 and 100 and the leaderboard will be updated.

!!!! SUBMISSION DEADLINE: 16:30 !!!!

Help on the day is available from these guys:



OUR TEAM

JOHNATHAN DUGGAN
johnathan.j.duggan@aib.ie

MIKE COTTER
mike.a.cotter@aib.ie

FIONNUALA DOHERTY
fionnuala.m.doherty@aib.ie

KEVIN MCTIERNAN
kevin.m.mctiernan@aib.ie

TOMMY MITCHELL
tommy.j.mitchell@aib.ie

JAMIE ROCHE
jamie.t.roche@aib.ie

EOIN MURPHY
eoin.g.murphy@aib.ie

PETER KOVAR
peter.x.kovar@aib.ie

AONGHUS COLLINS
aonghus.f.collins@aib.ie

Finally…..GOOD LUCK!

Good luck everyone. Enjoy the day, the pizza, the music, the air hockey, the big fuzzy blue thing??... and of course the data. May the best team win!