GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

Week 5 – Recommendation Engines
Tuesday 20th June 2017

AGENDA 2

1. Guest Speaker!
2. What are Recommendations? (You tell me)
3. What is the motivation of recommendations?
4. What is Content-Based Filtering?
5. What is Collaborative Filtering?
6. Measuring Accuracy
7. Lab
8. Other Considerations
9. Discussion

DATA SCIENCE PART TIME COURSE

WHAT ARE RECOMMENDATION ENGINES?

- What are recommendations?
- Why are they important?
- Give one example people are likely to come across?

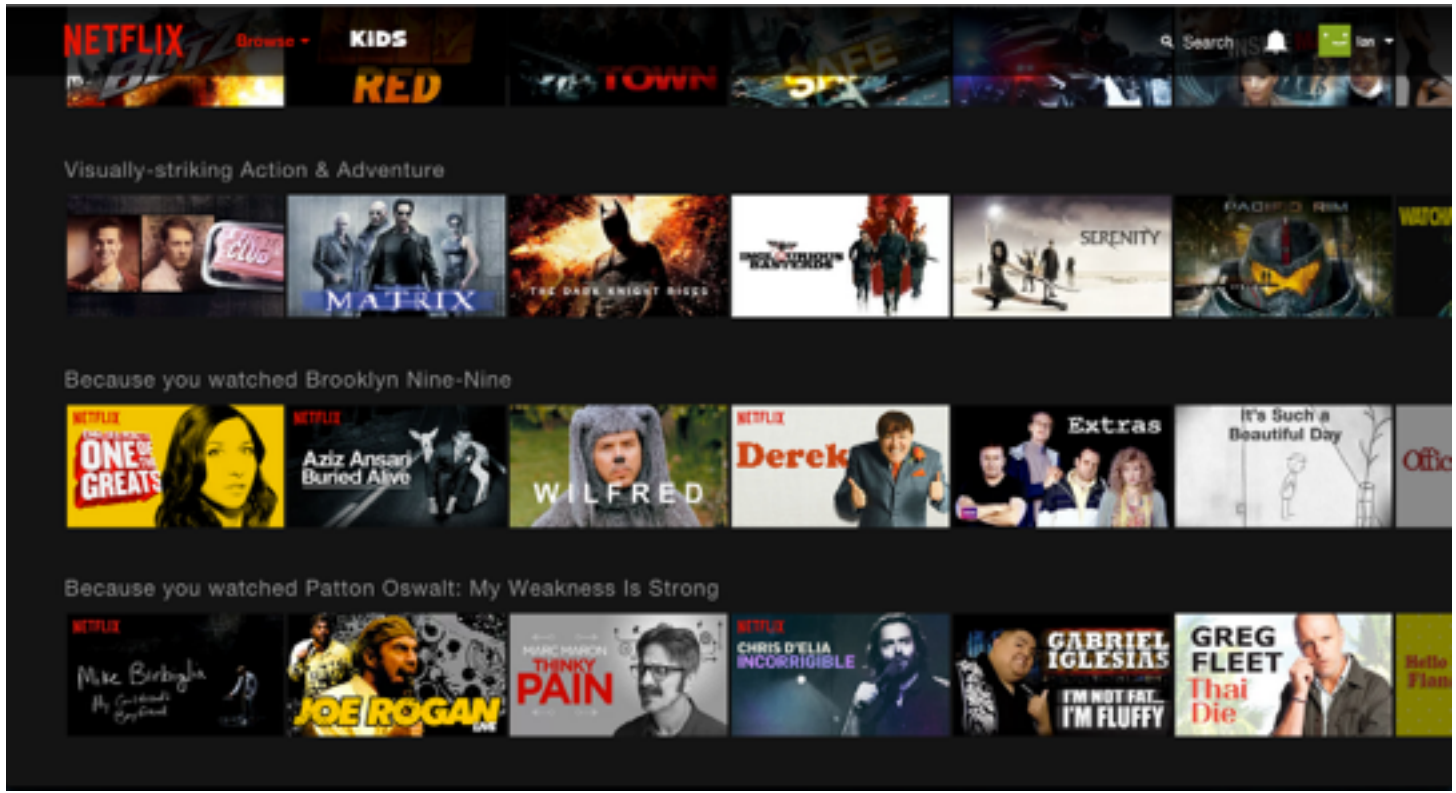
Work in two groups to answer the above questions and present back to the class.

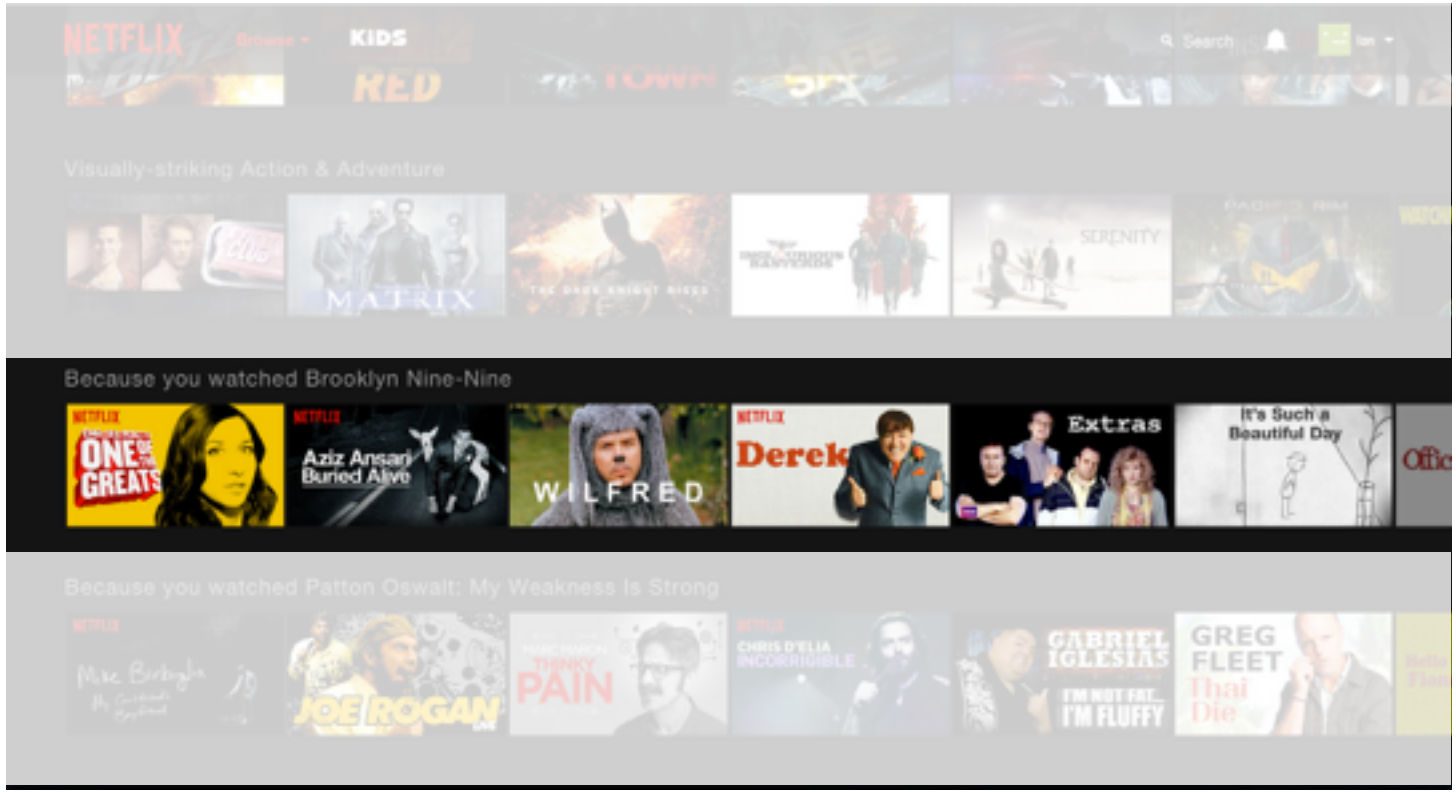
5 mins

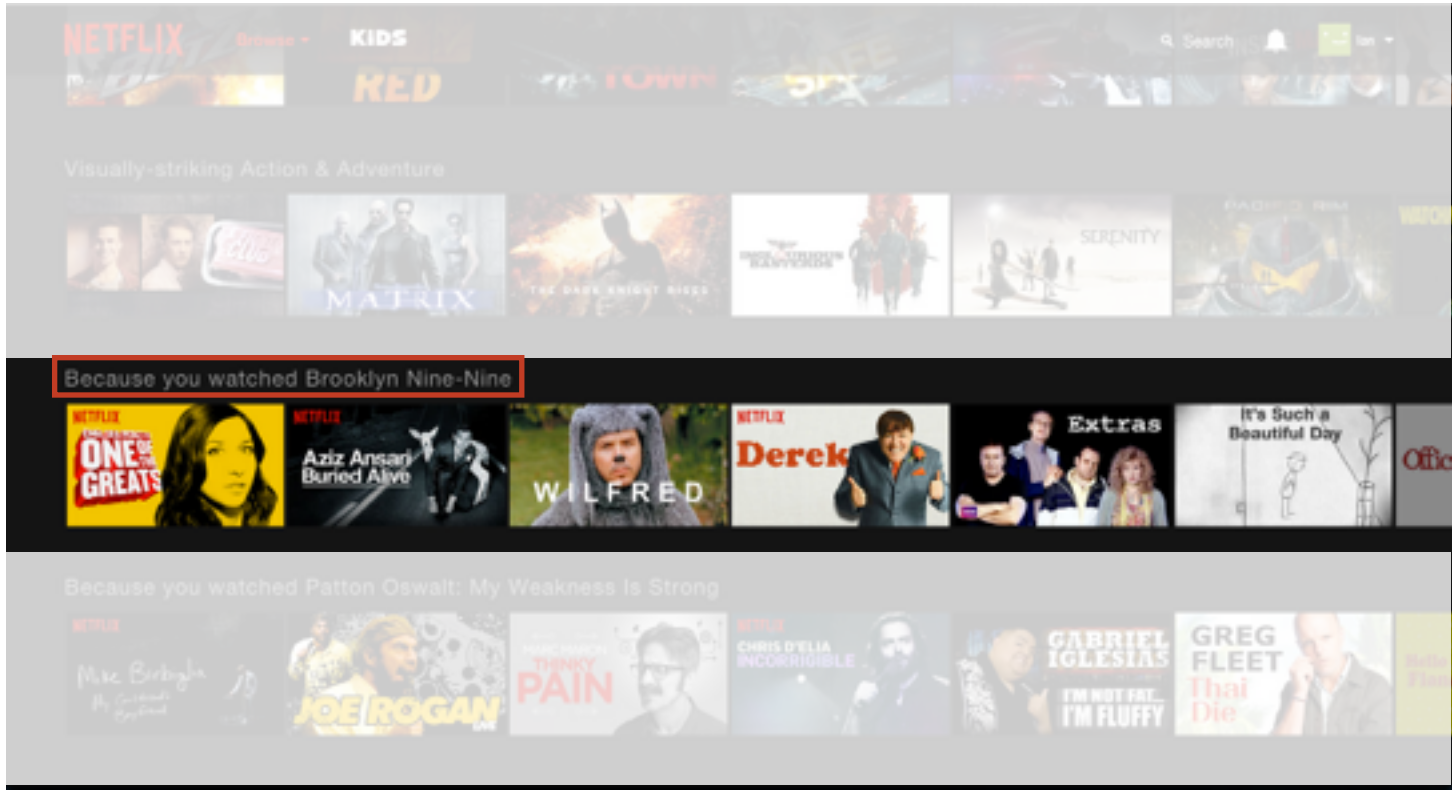
Recommendation engines aims to match users to things (movies, songs, items, events, etc) they might enjoy but have not yet tried.

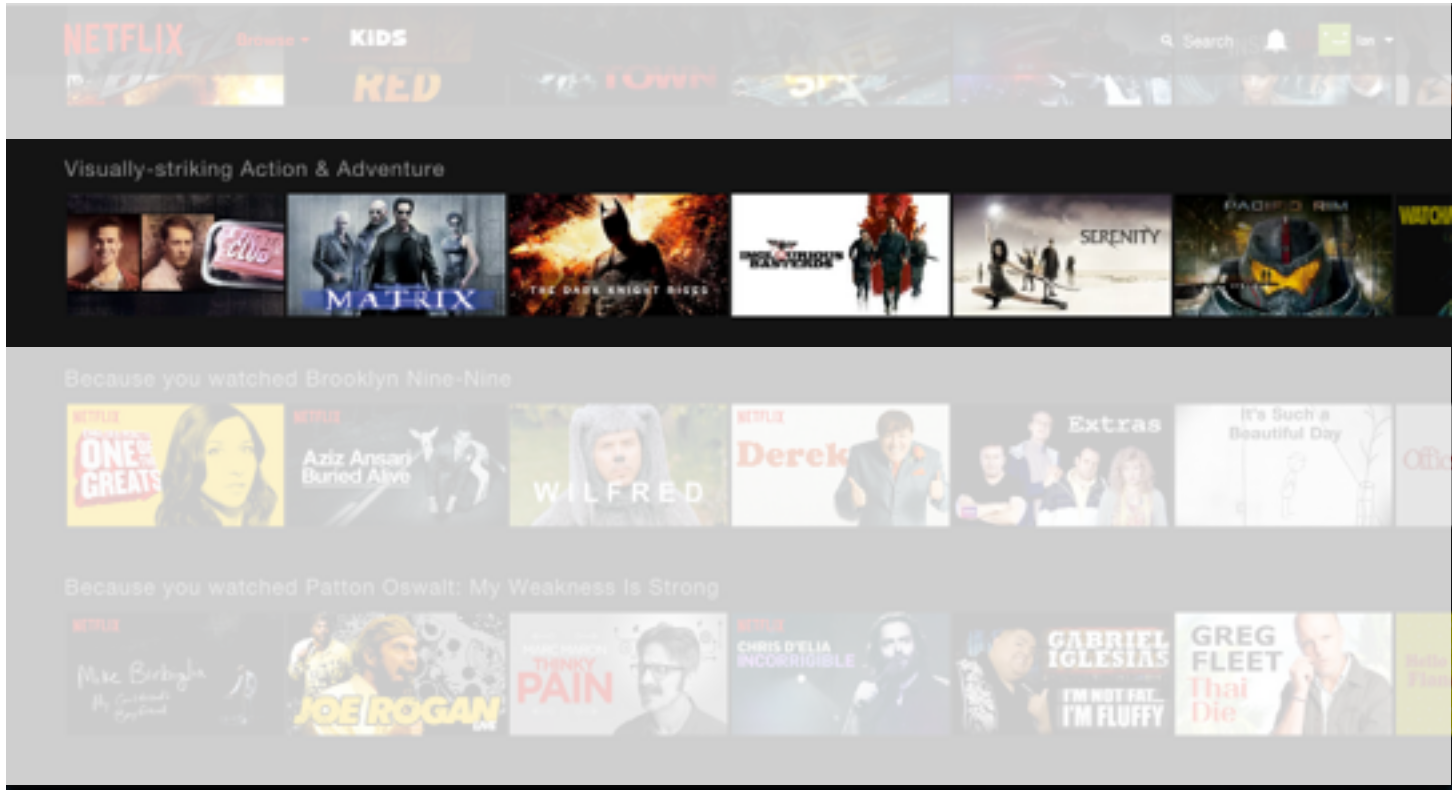
The rating is produced by analysing other user/item ratings (and sometimes item characteristics) to provide personalised recommendations to users.

DATA SCIENCE PART TIME COURSE


NETFLIX
Browse
KIDS
Search
Visually-striking Action & Adventure
MATRIX
SERENITY
Because you watched Brooklyn Nine-Nine
Aziz Ansari
Buried Alive
WILFRED
Derek
Extras
It's Such a
Beautiful Day
Because you watched Patton Oswalt: My Weakness Is Strong
JOE ROGAN
PAIN
CHRIS D'ELIA
INCORRIGIBLE
GABRIEL
IGLESIAS
I'M NOT FAT...
I'M FLUFFY
GREG
FLEET
Thai
Die


NETFLIX
Browse
KIDS
Search
Visually-striking Action & Adventure
MATRIX
SERENITY
Because you watched Brooklyn Nine-Nine
Aziz Ansari
Buried Alive
WILFRED
Derek
Extras
It's Such a
Beautiful Day
Because you watched Patton Oswalt: My Weakness Is Strong
JOE ROGAN
PAIN
CHRIS D'ELIA
INCORRIGIBLE
GABRIEL
IGLESIAS
I'M NOT FAT...
I'M FLUFFY
GREG
FLEET
Thai
Die


NETFLIX
KIDS
Visually-striking Action & Adventure
MATRIX
SERENITY
Because you watched Brooklyn Nine-Nine
Aziz Ansari
Buried Alive
WILFRED
Derek
Extras
It's Such a
Beautiful Day
Because you watched Patton Oswalt: My Weakness Is Strong
JOE ROGAN
PAIN
GABRIEL
IGLESIAS
I'M NOT FAT...
I'M FLUFFY
GREG
FLEET
Thai
Die


NETFLIX
Browse
KIDS
Search
Visually-striking Action & Adventure
MATRIX
SERENITY
Because you watched Brooklyn Nine-Nine
Aziz Ansari
Buried Alive
WILFRED
Extras
It's Such a Beautiful Day
Because you watched Patton Oswalt: My Weakness Is Strong
JOE ROGAN
PAIN
GABRIEL IGLESIAS
I'M NOT FAT... I'M FLUFFY
GREG FLEET
Thai Die


NETFLIX
2009
DATE 09.21.09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION
00/100
FOR The Netflix Prize
Reed Hastings

There are two general approaches to the design:

In **content-based** filtering, items are mapped into a feature space, and recommendations depend on item characteristics.

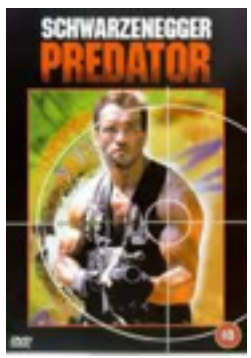
In contrast, the only data under consideration in **collaborative filtering** are user-item ratings, and recommendations depend on user preferences.

DATA SCIENCE PART TIME COURSE

CONTENT-BASED FILTERING

Looking at attributes of an item, you then make recommendations based on how similar those items are.

You liked Predator with Arnold Schwarzenegger you might also like The Terminator (because Arnie's in that too).



Content-based filtering begins by mapping each item into a feature space. Both users and items are represented by vectors in this space.

Item vectors measure the degree to which the item is described by each feature, and user vectors measure a user's preferences for each feature.

Ratings are generated by taking dot products of user & item vectors.

Pandora is an example of Content-Based filtering. A massive taxonomy of musical information. Trained musical analysts identify over 450 musical characteristics (lookup the Music Genome Project).

Content-based filtering has some difficulties:

- Must map items into a feature space (manual work)
- Recommendations are limited in scope (items must be similar to each other)
- Hard to create cross-content recommendations (eg books/music films...this would require comparing elements from different feature spaces)

DATA SCIENCE PART TIME COURSE

COLLABORATIVE FILTERING

“Customers who purchased X also purchased Y”

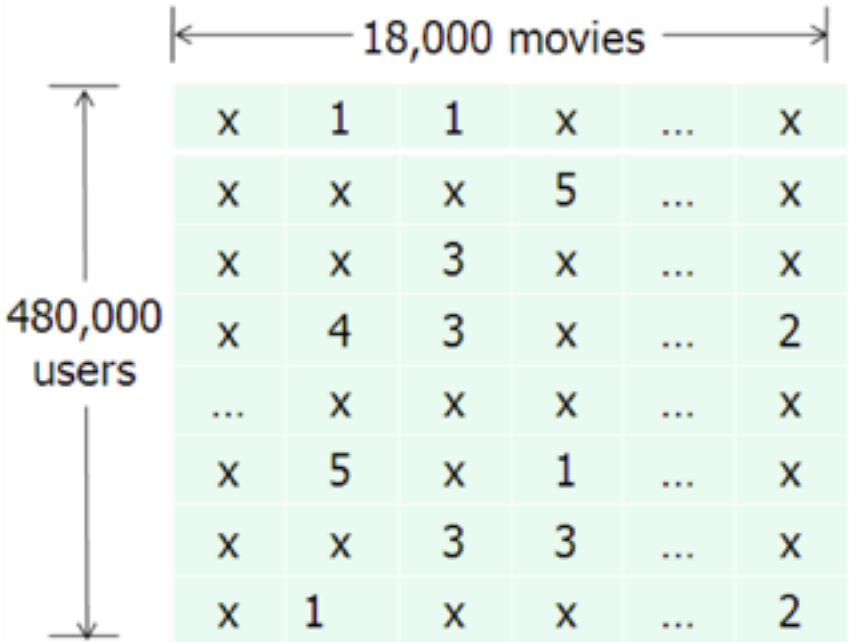
Someone with similar tastes to you will be able to recommend things you might like, e.g. people who watch ‘The Newsroom’ will probably enjoy ‘The Social Network’ because there is a large audience in common.

Collaborative filtering refers to a family of methods for predicting ratings where instead of thinking about users and items in terms of a feature space, we are only interested in the existing user-item ratings themselves.

In this case, our dataset is a ratings matrix whose columns correspond to items, and whose rows correspond to users.

This will be the general form of the data we analyse for collaborative filtering.

The method relies on previous user-item ratings (or feedback).



Collaborative filtering is susceptible to the Cold Start problem.

What happens if we don’t have any (or enough) reviews?

Collaborative filtering is susceptible to the Cold Start problem.

What happens if we don't have any (or enough) reviews?

Until users rate several items, we don't know anything about their preferences.

We can get around this by enhancing our recommendations using implicit feedback, which may include things like item browsing behaviour, search patterns, purchase history, etc. Or by using a hybrid model.

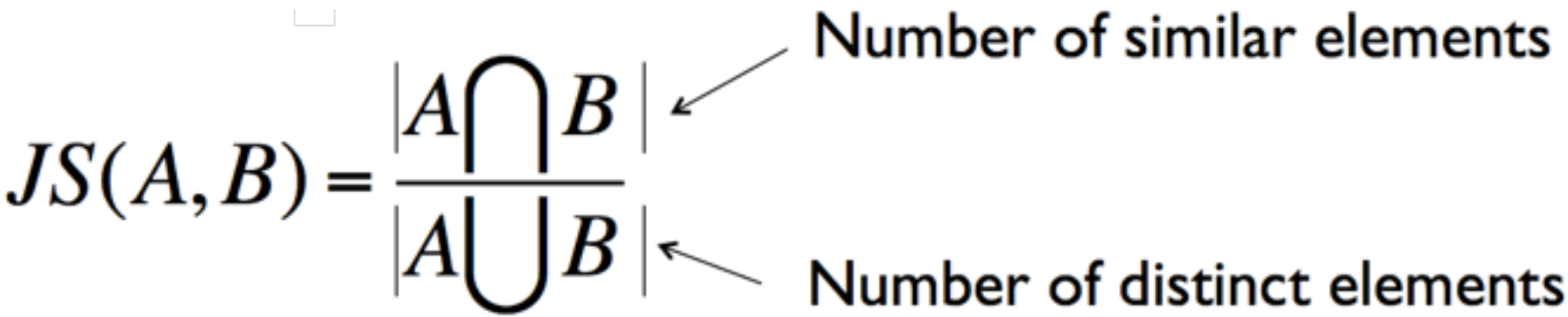
DATA SCIENCE PART TIME COURSE

SIMILARITY SCORES

Jaccard Similarity:
Defines similarity between two sets of objects

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

Jaccard Similarity:
Defines similarity between two sets of objects



$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

JS ({1, 2, 3}, {2,3,4}) = {2,3} / {1,2,3,4}
= 2/4
= 1/2

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

User one: {“LG-LCDTV 52” ”,“Blu Ray Player”,“HDMI Cable”}
User two: {“LG-LCDTV 52”,“Sony PS4”,“HDMI Cable”}

JS (User one, User two) =

$$JS(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

User one: {“LG-LCDTV 52” ”,“Blu Ray Player”,“HDMI Cable”}
User two: {“LG-LCDTV 52”,“Sony PS4”,“HDMI Cable”}

JS (User one, User two) = 2/4
= 1/2

LAB


CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
demotivateur.fr

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m "descriptive label for the commit"
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

OTHER CONSIDERATIONS

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	Positive	**c**	**d**

Accuracy = Correct Recommendations / Total Possible Recommendations
= (a + d)/(a + b +c + d)

Precision = Correctly Recommended Items / Total Recommended Items
= d / (b + d)

Recall = Correctly Recommended Items / Total Useful Recommended Items
= d / (c + d)

Precision ~1 means the algorithm returned more relevant results than irrelevant.
Recall ~1 means that an algorithm returned most of the relevant results.

Explicit data is when you ask the user to rate something, e.g. 1-5 star rating for a movie

Implicit data is when you observe a users behaviour and record

- Alternating Least Squares (ALS)
- Stochastic Gradient Descent (SGD)
- Singular Value Decomposition (SVD)
- Factorization Machine (FM)
- Collaborative Less is More Filtering (CLiMF)

- Ranking
- Freshness
- Diversity
- Social Recommendations
- Context Aware Recommendations
- Hybrid Models (Combining Content based filtering and Collaborative filtering)
- Model Objectives (what are you trying to optimise)
- Sequences?

DATA SCIENCE – Week 5 Day 1

DISCUSSION TIME

- Week 4 Review
- Homework 2
- Pre-reading
- Halfway Review
- Project

DISCUSSION TIME

- Regularisation
  - Why do we do this?
  - What are 3 types of regularised regression?
  - How does it work?
- Clustering
  - Why do we do this?
  - How does k-means work?

DATA SCIENCE – Week 5 Day 1

HOMEWORK 2

Overview

- Describe linear Regression
- Relate what we've learned in class to practical skills mentioned in a paper
- Work on your projects (Main Part)
- Due Friday 30th June 12 midnight

HALFWAY REVIEW

- Saturday July 1st 10 am – 3 pm
- Optional class reviewing the material in the first half then helping with projects in the second part
- Works best if you have specific questions so review the materials beforehand
- We are going to be in the smaller class room next door so first come first served

PRE-READING

- AirBnB article on AutoML: https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8