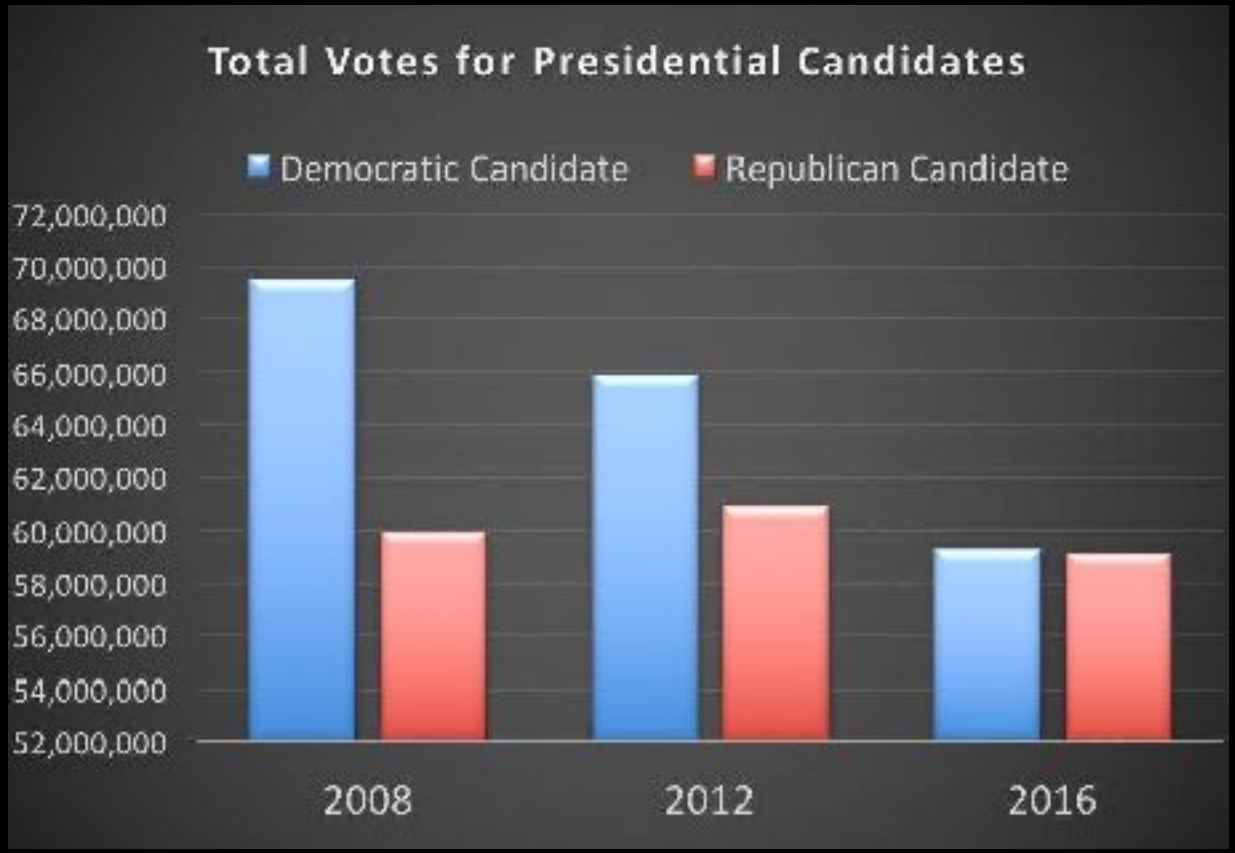
GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 6

Week 6 – Recommendation Engines
Monday 14th November

THOUGHTS ON GRAPHS



1. Mid-Course Surveys!
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. What matters for implementation
10. 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.

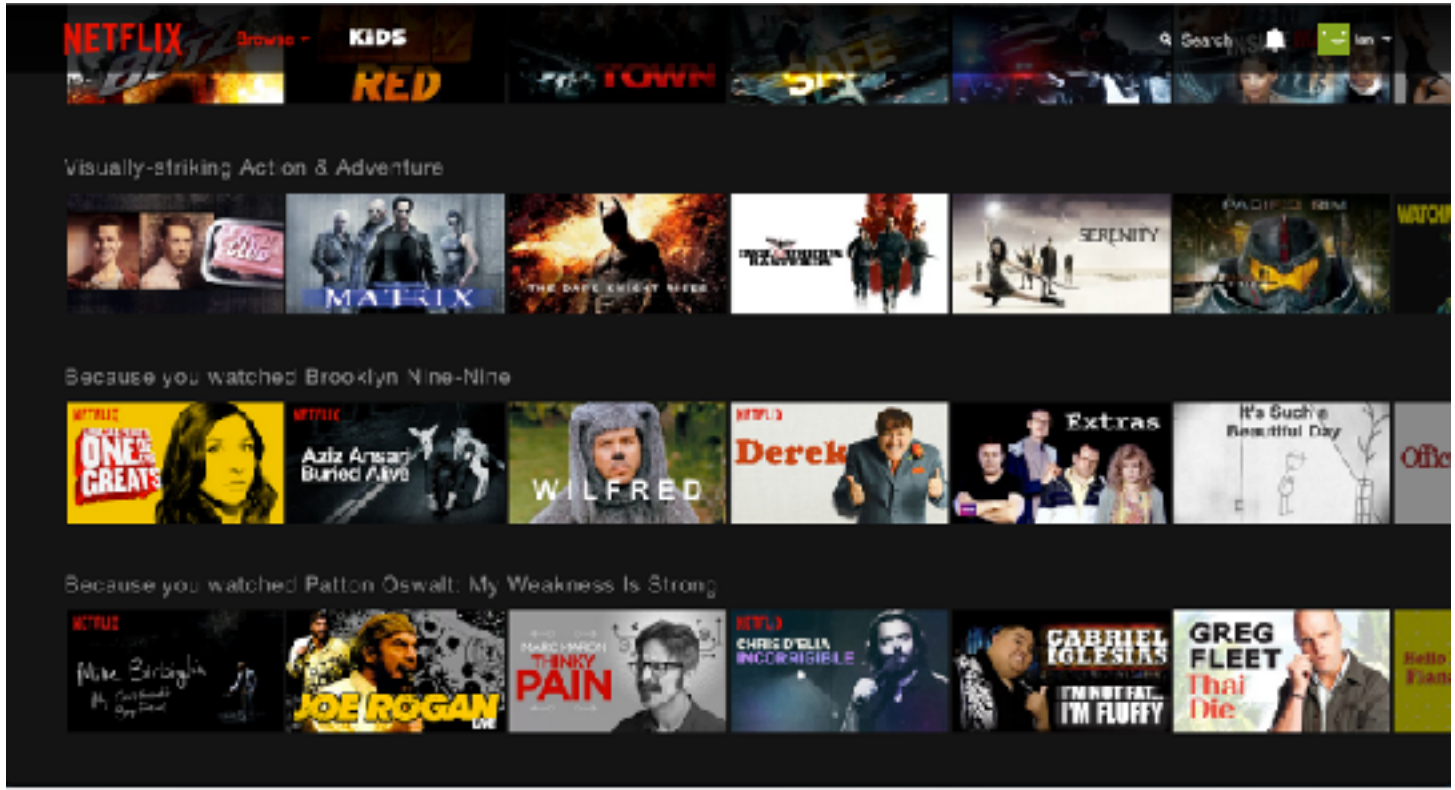
"The Web is leaving the era of search and entering one of discovery. What's the difference?

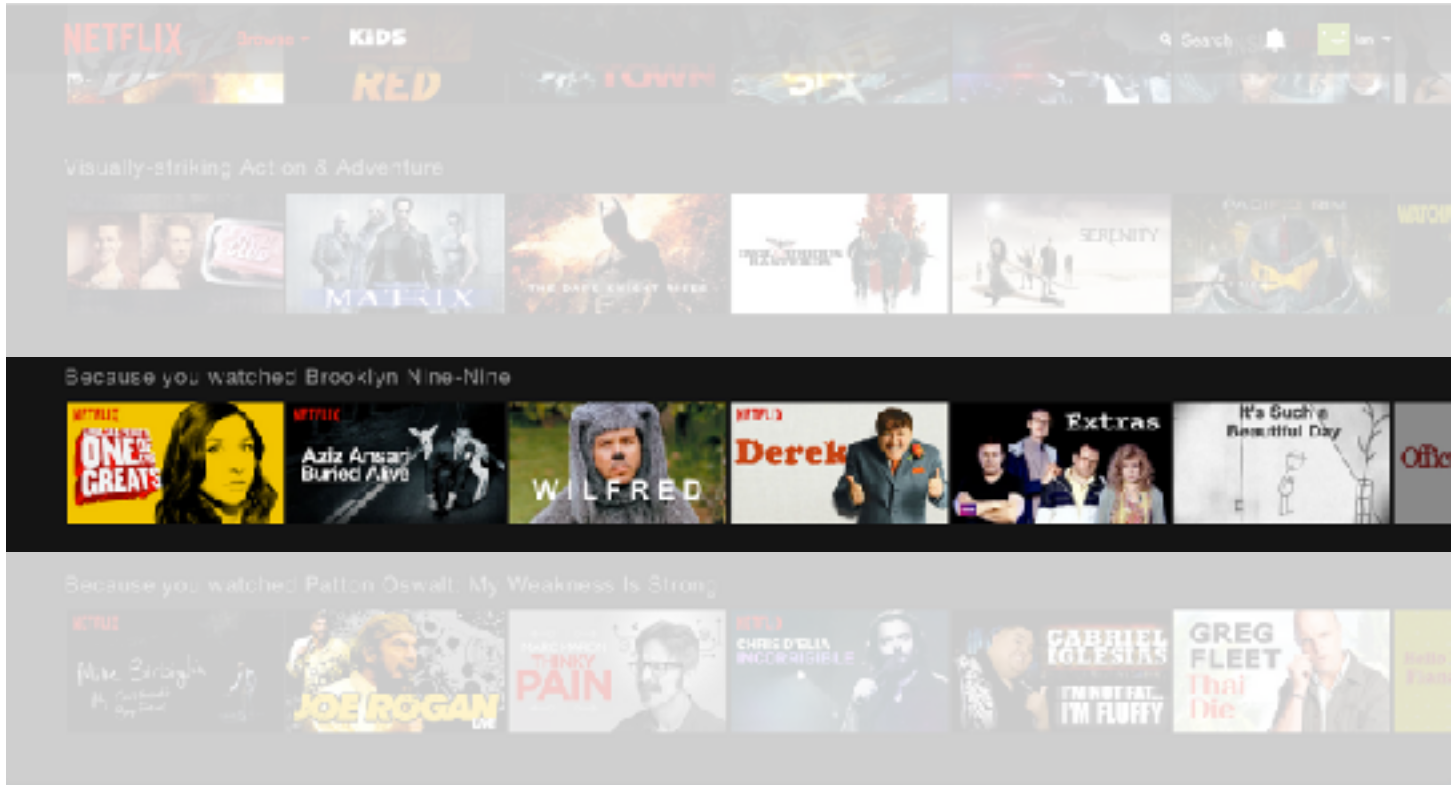
Search is what you do when you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you." – CNN Money, "The race to create a 'smart' Google

- Netflix: 2/3 of the movies watched are recommended
- Google News: recommendations generate 38% more click-throughs
- Amazon: 35% sales from recommendations
- Choicestream: 28% of the people would buy more music if they found what they liked.

Estimate a **utility function**
to **predict** how
a user will **like** an item.

DATA SCIENCE PART TIME COURSE


NETFLIX
Browse
KIDS
RED
TOWN
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 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
TINDY PAIN
GABRIEL IGLESIAS
I'M FLUFFY
GREG FLEET
That Die


NETFLIX
2008
DATE 04.21.09
BellMor's Pragmatic Chaos
$1,000,000.00
ONE MILLION
00/100
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.

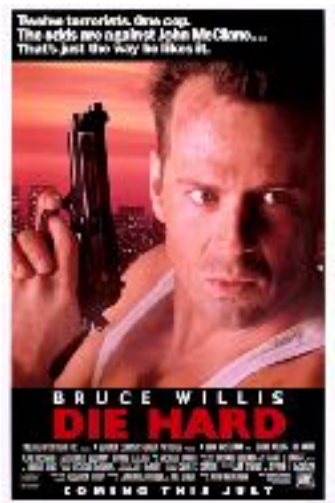
Item vectors measure the degree to which the item is described by each feature, and user vectors measure a user's preferences for each feature.

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

Recommendations are based on the information on the content of items rather than content of items on other users' opinions.

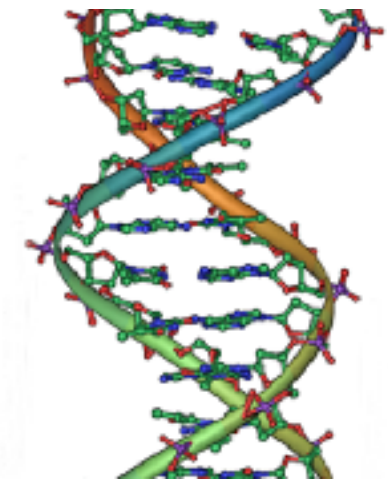
Use a machine learning algorithm to model the users' preferences from examples based on a description of the content.

Explicit attributes or characteristics
e.g. for a movie:
Genre: Action / adventure
Feature: Bruce Willis
Year: 1995



- Suitable for text-based products (web pages, books) Items are "described" by their features (e.g. keywords)
- Users are described by the keywords in the items they bought
- Recommendations based on the match between the content (item keywords) and user keywords
- The user model can also be a classifier (Neural Networks, SVM, Naïve Bayes...)

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



PANDORA

Content-based filtering has some difficulties:

- Must map items into a feature space (potentially 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).

Each user has expressed an opinion for some items:

- Explicit opinion: rating score
- Implicit: purchase records or listen to tracks

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

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Number of similar elements (intersection)

Number of distinct elements (union)

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

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: {“Target”,“Banana Republic”,“Old Navy”}
User two: {“Banana Republic”,“Gap”,“Kohl’s”}

JS (User one, User two) =

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

User one: {“Target”,“Banana Republic”,“Old Navy”}
User two: {“Banana Republic”,“Gap”,“Kohl’s”}

JS (User one, User two) = 1/5

LAB

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 PART TIME COURSE

WHAT MATTERS FOR IMPLEMENTATION OF A RECOMMENDATION SYSTEM

Exploration

Diversity

Speed

Not the last fraction of a percent

Deployable

– Clever prototypes don’t count

Robust

– Mishandling is common

Transparent

– Will degradation be obvious?

Proportionate

– Where is the highest value per minute of effort?

What are the most important algorithmic advances in recommendations over the last 10 years?

1. Result dithering
2. Anti-flood

Dithering is used to re-order recommendation results

Re-ordering is done randomly

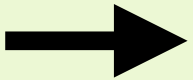
Dithering is guaranteed to make off-line performance worse

Dithering also has a near perfect record of making actual performance much better

"Made more difference than any other change"

error = 0.5

1	2	6	5	3	4	13	16
1	2	3	8	5	7	6	34
1	4	3	2	6	7	11	10
1	2	4	3	15	7	13	19
1	6	2	3	4	16	9	5
1	2	3	5	24	7	17	13
1	2	3	4	6	12	5	14
2	1	3	5	7	6	4	17
4	1	2	7	3	9	8	5
2	1	5	3	4	7	13	6
3	1	5	4	2	7	8	6
2	1	3	4	7	12	17	16



error = 0.7

1	2	8	3	9	15	7	6
1	8	14	15	3	2	22	10
1	3	8	2	10	5	7	4
1	2	10	7	3	8	6	14
1	5	33	15	2	9	11	29
1	2	7	3	5	4	19	6
1	3	5	23	9	7	4	2
2	4	11	8	3	1	44	9
2	3	1	4	6	7	8	33
3	4	1	2	10	11	15	14
11	1	2	4	5	7	3	14
1	8	7	3	22	11	2	33

Recommendations for soap:

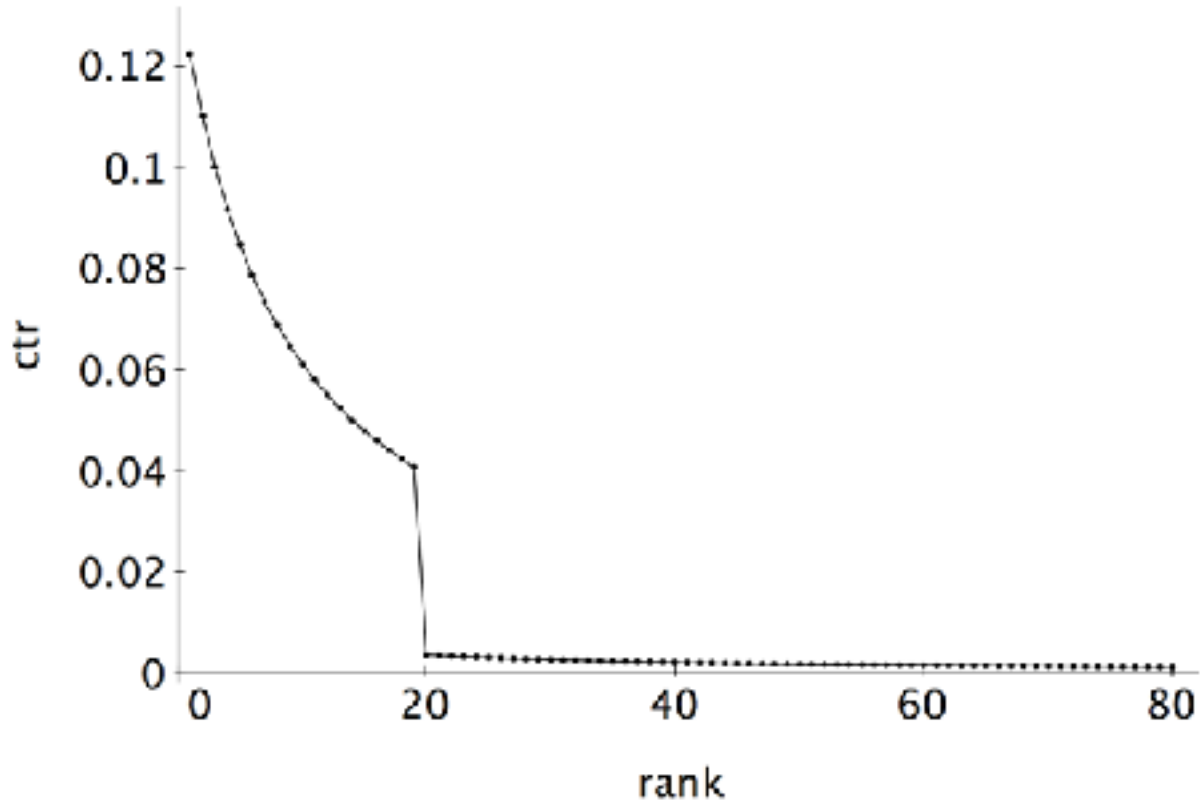
1. Pears soap
2. Pears soap sensitive
3. Pears soap for men
4. Pears soap for children
5. J&J Soap

Recommendation systems will give endless variations on the same thing unless prevented

Solution:

Penalise the rank for any result that appears too similar to higher-ranked results.

1. Pears soap
2. J&J Soap
3. Dolphin friendly dish detergent
4. Sugar soap
5. Floor Cleaner


0.12
0.1
0.08
0.06
0.04
0.02
0
ctr
0
20
40
60
80
rank

HOMEWORK

Homework

- Homework 2 – Due Monday 14th of November
- Sign up for Databricks community edition https://databricks.com/try-databricks

Read the following

- Research about Spark