ICS 4U - Team Project

Books, Ratings, Suggestions and More!

Note to Printed Slides

If you are reading this only on the materials provided by the Summer Institute and did not see the actual presentation, you may not be able to understand the slides on their own. They describe an assignment idea for ICS4U created by a team of CS educators including high school and university teachers. You can obtain the editable assignment handout and a teachers' guide by contacting Michelle Craig at mcraig@cs.toronto.edu

Netflix Contest

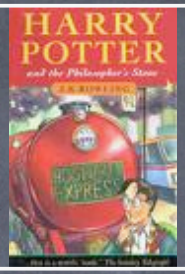
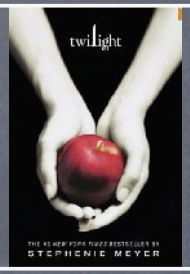
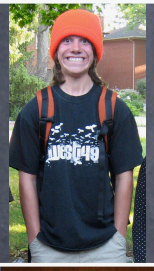
- movie rental business
- want to recommend movies to their customers based on the previous ratings of those same customers
- 1 Million $$ prize
- 51051 contestants, 41305 teams, 186 countries
- 3 years - closed this summer

chapters.indigo.ca

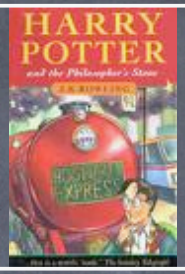
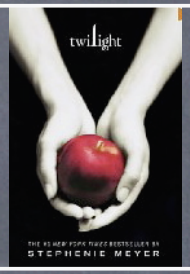
- Customers Who Bought This Item Also Bought ...
- May We Also Recommend ...

Overview

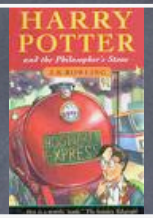
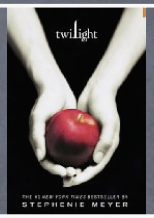
- a collection of items (books) probably a fixed set
- an initial collection of customers
- initial customers have rated all the items
- a new customer
  - rates some items
  - receives recommendations on others

					
	★	★		X	★
		--	★	X	★
	★		X	★	--

Michelle Craig, University of Toronto, 2009

					
	5	3	0	-3	5
	0	1	5	-5	3
	3	0	-3	5	1

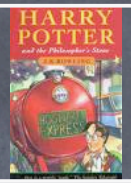
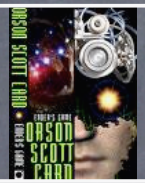
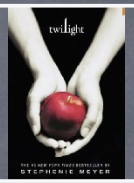
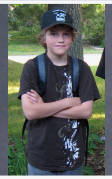
Michelle Craig, University of Toronto, 2009

					
	5	3	0	-3	5
	0	1	5	-5	3
	3	0	-3	5	1
new customer	5			1	3

Michelle Craig, University of Toronto, 2009

Approach A: Basic

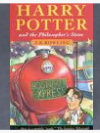
- Don't use the ratings provided by the new user
- Except to see if they have already read a book or not
- Use the average ratings from everyone else to put the books in order
- Recommend the most-highly rated books that this user hasn't yet read

					
	5	3	0	-3	5
	0	1	5	-5	3
	3	0	-3	5	1
→ AVG	4	2	1	-1	3
new customer	5			1	3

Michelle Craig, University of Toronto, 2009

Approach B

- Find the customer already in our database who is most similar to the new customer
- Use only the ratings from the most similar customer to make recommendations

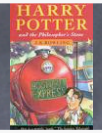
					
	5	3	0	-3	5
	0	1	5	-5	3
	3	0	-3	5	1
new	5	0	0	1	3

$5 \times 5 + 3 \times 0 + 0 \times 0 + -3 \times 1 + 5 \times 3 = 37$

$0 \times 5 + 1 \times 0 + 5 \times 0 + -5 \times 1 + 3 \times 3 = 4$

$3 \times 5 + 0 \times 0 + -3 \times 0 + 5 \times 1 + 1 \times 3 = 23$

Michelle Craig, University of Toronto, 2009

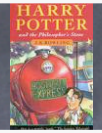
						
	5	3	0	-3	5	37
	0	1	5	-5	3	4
	3	0	-3	5	1	23
new	5	0	0	1	3	

$5 \times 5 + 3 \times 0 + 0 \times 0 + -3 \times 1 + 5 \times 3 = 37$

$0 \times 5 + 1 \times 0 + 5 \times 0 + -5 \times 1 + 3 \times 3 = 4$

$3 \times 5 + 0 \times 0 + -3 \times 0 + 5 \times 1 + 1 \times 3 = 23$

Michelle Craig, University of Toronto, 2009

						
	5	3	0	-3	5	37
	0	1	5	-5	3	4
	3	0	-3	5	1	23
new	5	0	0	1	3	

$5 \times 5 + 3 \times 0 + 0 \times 0 + -3 \times 1 + 5 \times 3 = 37$

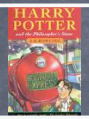
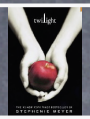
$0 \times 5 + 1 \times 0 + 5 \times 0 + -5 \times 1 + 3 \times 3 = 4$

$3 \times 5 + 0 \times 0 + -3 \times 0 + 5 \times 1 + 1 \times 3 = 23$

Michelle Craig, University of Toronto, 2009

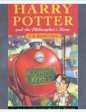
Approach C

Combine the rankings of lots of other customers

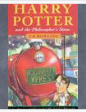
						
	5	3	0	-3	5	37
	0	1	5	-5	3	4
	3	0	-3	5	1	23
new	5			1	3	

Michelle Craig, University of Toronto, 2009

Approach C

					
	5x37	3x37	0x37	-3x37	5x37
	0x4	1x4	5x4	-5x4	3x4
	3x23	0x23	-3x23	5x23	1x23
Prediction					
new	5			1	3

Approach C

					
	5x37	3x37	0x37	-3x37	5x37
	0x4	1x4	5x4	-5x4	3x4
	3x23	0x23	-3x23	5x23	1x23
Prediction	4,324	115	-49	-16	220
new	5			1	3

Learning Expectations?

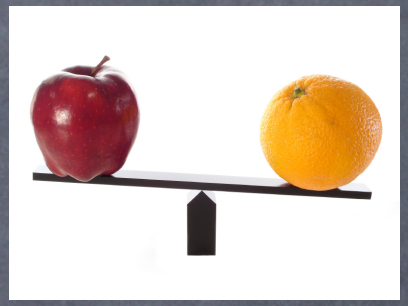
- A1.5 1D arrays of compound data types
- A2.1 modular program that is divided into multiple classes or files
- A2.2 modular design concepts
- A3.1 read/write to external file
- A3.4 sorting algorithms
- A3.5 algorithms processing 2D arrays

LE's continued ...

- A4.4 external user documentation
- B1 project management
- B2 software project expectations
- C1 modular design
- C2.3 compare efficiency in sorting algorithms
- D4 exploring computer science
- D2.1 investigate and analyse an ethical issue

Where is the Computer Science to highlight?

- Concept of a similarity measure between 2 items of the same type



- Distance measure
- In this project, we needed a distance measure between 2 customers and we used the dot product of their rating vectors.

More CS ...

- Once we have a distance measure, we can say how similar one instance is to any other instance in the universe.
- Think about the customers. Once we know how similar one particular customer is to every other customer, we can find the X most similar and do things like recommend friends.

Weighted averages

- In approach C, we calculate the final prediction values as a weighted average of the ratings from the other users.
- Weighted averages (where some contributions count more than others) come up all the time in computer science:
  - page-ranking algorithms
  - calculating complexity in searching/sorting

Sparse Data

- Would these 3 approaches work on the real Chapters/Indigo data?
  - 900,000 books purchased but less than 300,000 ratings
  - 25,000 customers who rated at least 1 book but on average each book rated by 3 customers
- Most of the data file is zeros!!!

Sparse Data

- Huge datasets but mostly empty
- Our algorithms wouldn't work
- That's why the NETFLIX prize was a challenge!
- Machine Learning techniques

Machine Learning - 101

- Use existing data set (called training data)
- Keep some data out (called test set)
- Learn a transformation that will generate (from the ratings set) a small representation for each book and each customer. The representations for book b and customer c must together be able to produce a good estimate of customer c's rating of book b.

Machine Learning

- Start with a random transformation and keep improving it by small changes that make it do a better and better job of getting the predictions right on the training data.
- Each time also test on the test data but don't use it for training. Stop when you start to get worse on the test data. That means that you are overfitting the peculiarities of the training data.

Computers and Society Topics

- An opportunity to talk about the power of large tasks accomplished by many tiny contributions (Web 2.0)
  - ratings of all sorts
  - Google maps - getting people to catalogue the globe in pictures
  - using CAPTCHA to help with OCR
  - www.hunch.com

Other Ethical Issues

- Some people have questioned the idea of Netflix using a contest to improve their core business model and wondered if we should encourage this as a society. Should programmers work on contests for free?
- What about open-source software? Is it trustworthy? Does it put programmers out of work?
- Digital profiling: How much personal data is out there about you?

Contact

Michelle Craig
mcraig@cs.toronto.edu
University of Toronto

www.cs.toronto.edu/ICS4U/BookRatingProject

Thanks to colleagues at TDSB for collaboration on this project (Karen, Janice, Myra, Ryk)