Collectieve Intelligentie

Collaborative Filtering

Huishoudelijk

- Correlatie tussen tijd op de dag en aanwezigheid

- Correlatie tussen aanwezigheid en cijfer

Gastcolleges & Project

- In de tweede helft zijn er veel gastcolleges vanuit allerlei hoeken en verschillende bedrijven

- Deze colleges zijn er o.a. om je inspiratie te geven voor het project

Tentamen

- Woensdag 9 uur op IWO (zie datanose)

- Op papier:
  - Geen code schrijven

- Stof:
  - Hoorcolleges
  - Opdrachten

Deze week

- Een recommender system maken

- Films aanraden aan gebruikers

Als je iemand die je niet kent een film wil aanraden, wat vraag je dan?

Welke genres vind je leuk?

Wat zijn je favoriete acteurs?

Wat zijn je favoriete regisseurs?

Hou je van documentaires?

Kijk je liever art-house of blockbuster films?

Welke andere films vind je goed?

Welke andere films vind je niet goed?

Welke andere films vind je goed?

Welke andere films vind je niet goed?

Experiment

Wie hebben de volgende films gezien?

Frozen Inception

Experiment

Wie hebben de volgende films gezien?

Frozen Inception

Wie hebben *bovendien* 2 of meer van de 3 onderstaande films gezien?

Titanic The Avengers Avatar

Experiment

Instructies:

1) **Pak een post it en schrijf je naam er op**

Experiment

Instructies:

1) Pak een post it en schrijf je naam op de voorkant
2) **Geef (in gedachte) een waardering tussen de 1 en de 5 voor ieder van de films:**

Frozen, Inception, Titanic, The Avengers en Avatar

Experiment

Instructies:

1) Pak een post it en schrijf je naam op de voorkant
2) Geef (in gedachte) een waardering voor de films
3) **Schrijf de waardering voor Frozen en Inception op de voorkant van de post-it**

Experiment

Instructies:

1) Pak een post it en schrijf je naam op de voorkant
2) Geef (in gedachte) een waardering voor de films
3) Schrijf de waardering voor Frozen en Inception op de voorkant.
4) **Schrijf de waardering voor Titanic, The Avengers en Avatar op de achterkant van de post-it**

Experiment

Instructies:

1) Pak een post it en schrijf je naam op de voorkant
2) Geef (in gedachte) een waardering voor de films
3) Schrijf de waardering voor Frozen en Inception op de voorkant.
4) Schrijf de waardering voor Titanic, The Avengers en Avatar op de achterkant
5) **Plak de post-it op de juiste locatie in de grafiek**

Experiment

Instructies:

1) Pak een post it en schrijf je naam op de voorkant
2) Geef (in gedachte) een waardering voor de films
3) Schrijf de waardering voor Frozen en Inception op de voorkant.
4) Schrijf de waardering voor Titanic, The Avengers en Avatar op de achterkant
5) Plak de post-it op de juiste locatie in de grafiek

User-based collaborative filtering

Doel:

- Raad een film aan

Strategie:

- Vind gebruikers die op jou lijken
- Gebruik de ratings van deze gebruikers om een predictie te maken

"Users like you also like Y"

Nu iets preciezer

We konden intuïtief zien dat gebruikers wel of niet op elkaar lijken.

Hoe zouden we dat iets iets preciezer kunnen maken?

Hoe laat je een computer dit doen?

Features

- Karakteristieke kenmerken

- Bij user-based collaborative filtering, interacties:
  - Ratings
  - Clicks
  - Aankopen
  - Bezocht
  - ...

Type Features

- Continu
  - 0.0 - 1.0
- Discreet
  - 1 2 3 4 5
- Nominaal / categorisch
  - man / vrouw
- Ordinaal
  - high medium low
- Interval
  - 1 - 100, 101 - 200, 201 - 300, 301 - 400

Feature Space

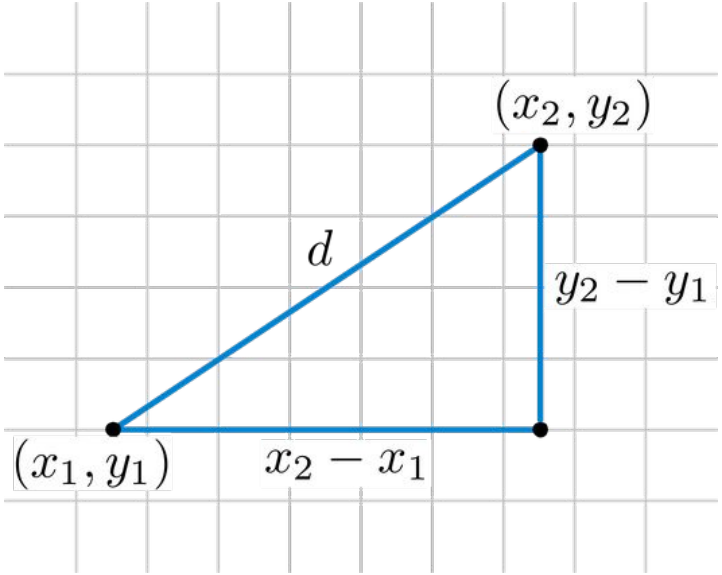
- Elke dimensie is een feature

- Stelt je in staat om gebruikers/items als een punt in de ruimte zien
  - Nu kunnen we afstand berekenen

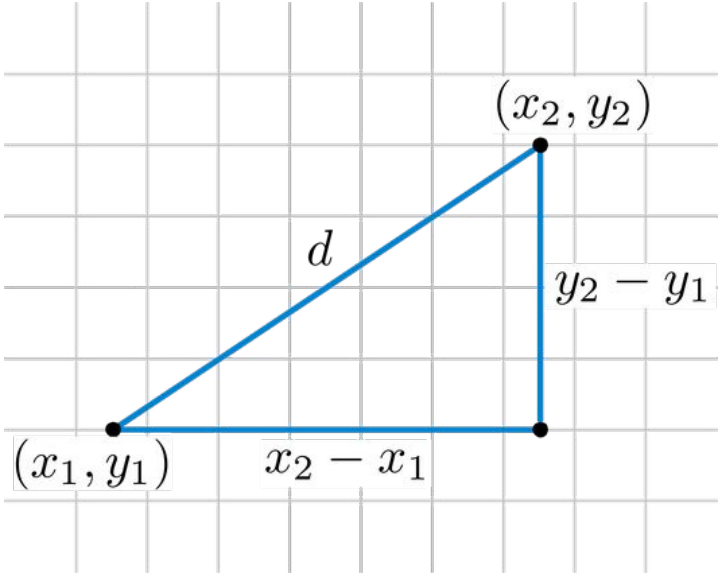
Utility Matrix

	Feature 1	Feature 2	Feature 3	...
User 1	3	4	5	
User 2	5	1	3	
User 3	4	4	5	
...				

Euclidische afstand



Euclidische afstand



$$distance = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

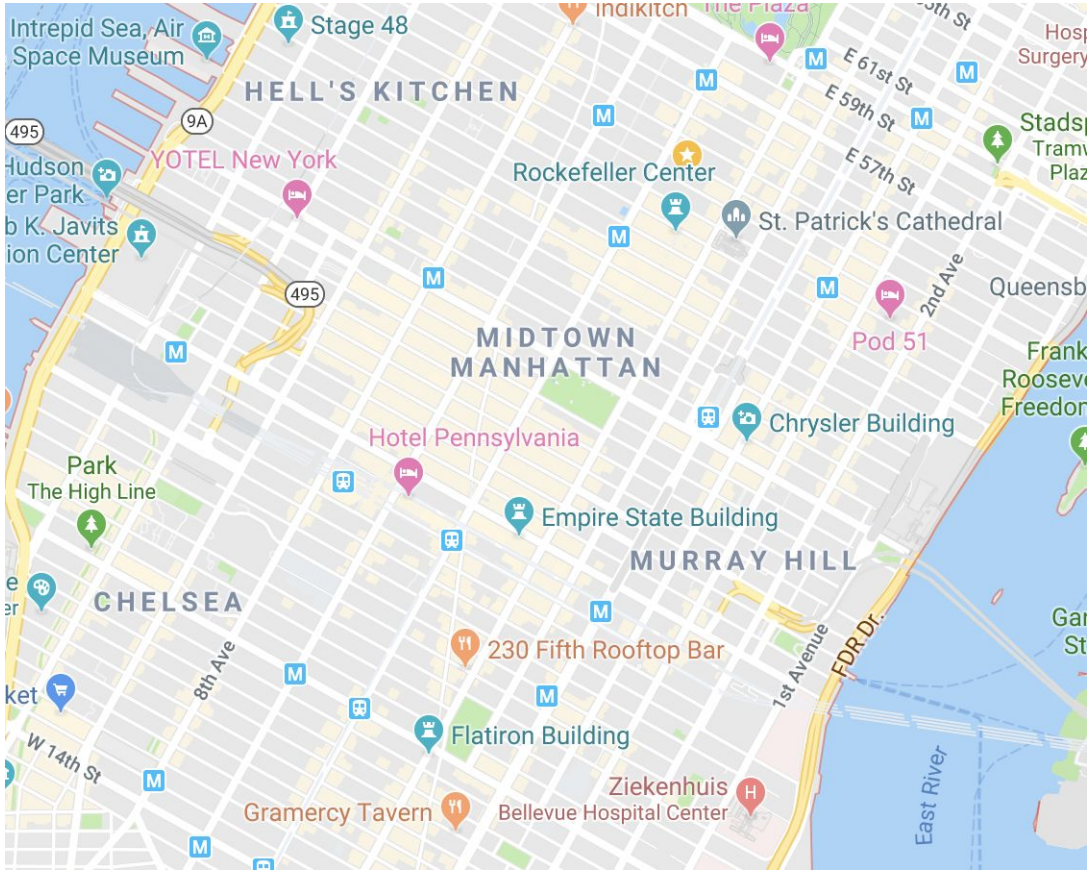
Similarity

$$similarity = 1/(1 + distance)$$

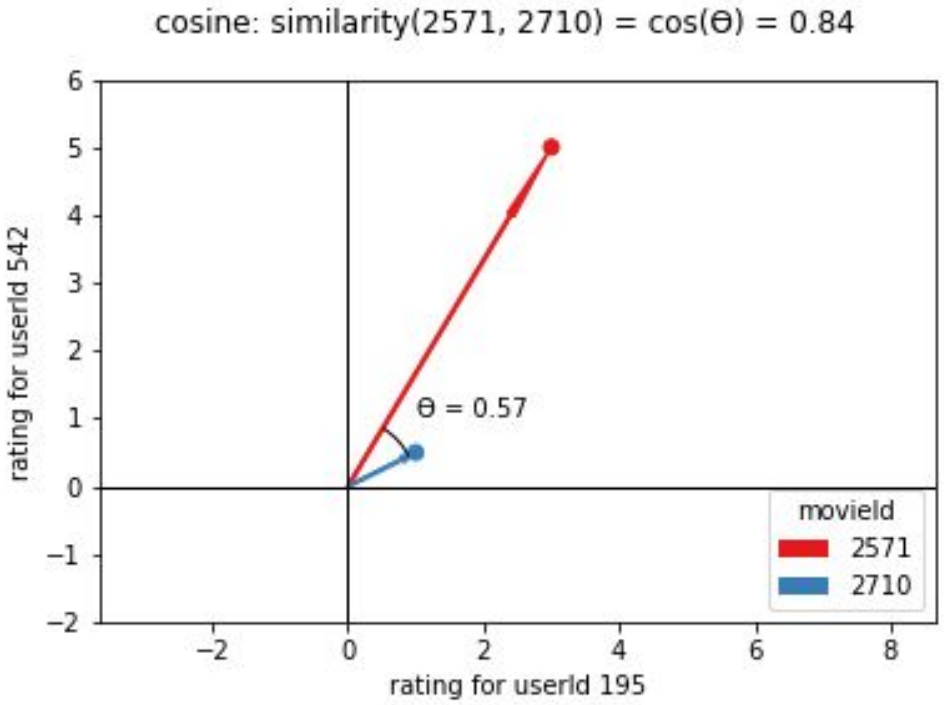
Meerdere similarity maten

- Manhattan distance
- Cosine similarity
- Pearson Correlation Coefficient

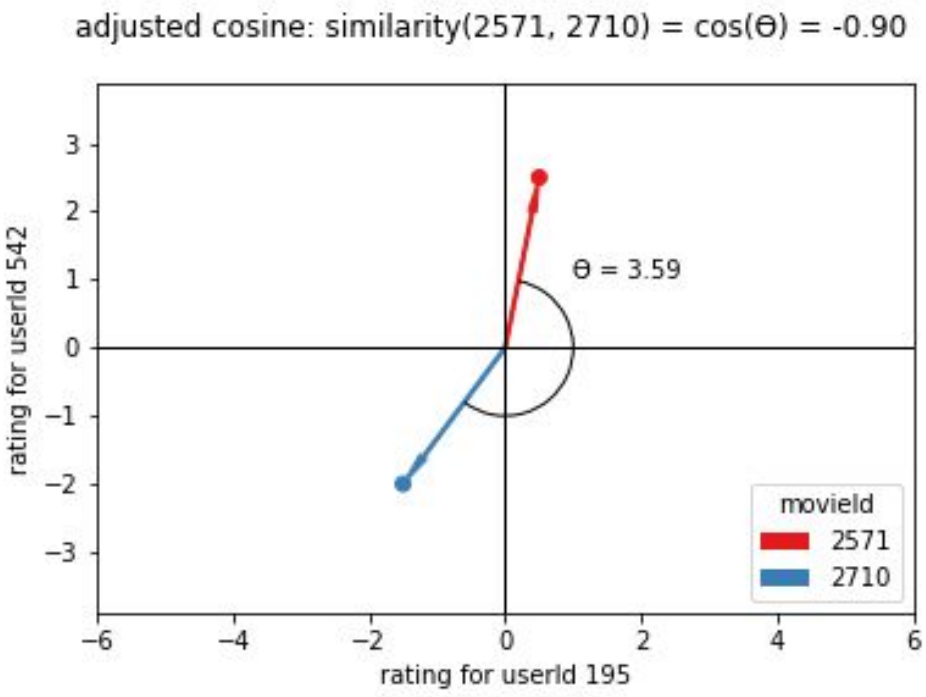
Manhattan distance



Cosine similarity



Adjusted cosine similarity



Similarity Matrix

	User 1	User 2	User 3	...
User 1	1.00	0.20	0.50	
User 2	0.20	1.00	0.21	
User 3	0.50	0.21	1.00	
...				

Hoe beveel je aan?

- Nearest Neighbor
  - Vraag de dichtstbijzijnde buur voor advies

- K Nearest Neighbor
  - Vraag de K dichtstbijzijnde buren voor advies

- Neighborhood
  - Vraag iedereen binnen een bepaalde straal voor advies

A nearest neighbor user-based collaborative filtering recommender system for movies

A nearest neighbor user-based collaborative filtering recommender system for movies

Probleem:

- Gegeven een film A die je nog niet hebt gezien, zou je film A willen zien?

A nearest neighbor user-based collaborative filtering recommender system for movies

Probleem:

- Gegeven een film A die je nog niet hebt gezien, zou je film A willen zien?

Aanpak:

1. Zet alle ratings en gebruikers in een utility matrix

A nearest neighbor user-based collaborative filtering recommender system for movies

Probleem:

- Gegeven een film A die je nog niet hebt gezien, zou je film A willen zien?

Aanpak:

1. Zet alle ratings en gebruikers in een utility matrix
2. Bouw een similarity matrix tussen gebruikers
    a. Bereken de afstand tussen elke gebruiker (euclidean distance)
    b. Bereken de similarity afhankelijk van de afstand

A nearest neighbor user-based collaborative filtering recommender system for movies

Probleem:

- Gegeven een film A die je nog niet hebt gezien, zou je film A willen zien?

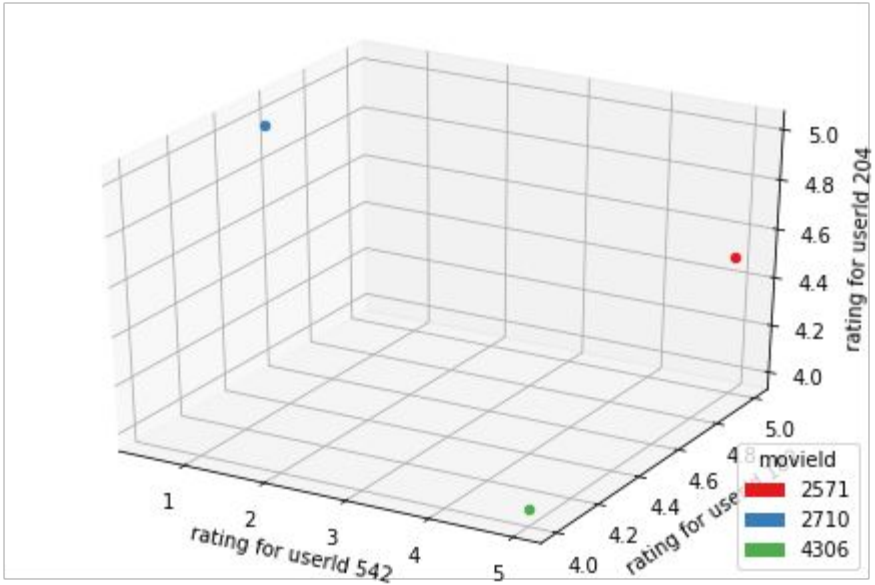
Aanpak:

1. Zet alle ratings en gebruikers in een utility matrix
2. Bouw een similarity matrix tussen gebruikers
    a. Bereken de afstand tussen elke gebruiker (euclidean distance)
    b. Bereken de similarity afhankelijk van de afstand
3. Gebruik de rating van de meest similar gebruiker die film A heeft ge-rate

Opschalen

- Meer features
  - Feature space wordt hoger dimensionaal
  - Lastig visueel voor te stellen
  - Similarity maten schalen makkelijk op

- Meer items
  - Meer punten in de feature space
  - Rekenkundig zwaar



Problemen

- Weinig ratings
- Similarities tussen alle gebruikers vinden is computationeel duur
- Gebruikers veranderen snel

Item-based collaborative filtering

Doel:

- Raad een film aan

Strategie:

- Raad films aan die lijken op films die jij leuk vindt

"Users who like X also like Y"

Utility Matrix

	Feature 1	Feature 2	Feature 3	...
~~User 1~~ Item 1	3	4	5	
~~User 2~~ Item 2	5	1	3	
~~User 3~~ Item 3	4	4	5	
...				

Similarity Matrix

	~~User 1~~ Item 1	~~User 2~~ Item 2	~~User 3~~ Item 3	...
~~User 1~~ Item 1	1.00	0.20	0.50	
~~User 2~~ Item 2	0.20	1.00	0.21	
~~User 3~~ Item 3	0.50	0.21	1.00	
...				

Collaborative filtering

Doel: raad een film aan

User-based filtering: Vind gebruikers die op elkaar lijken

Item-based filtering: Vind films die op elkaar lijken

Evaluatie

- Welke aanpak, item- of user-based?
- Welke features?
- Welke similarity maat?

Testen

- Voorspellingen werken enkel op “nieuwe data”

- Maar van nieuwe data weten we niet wat een correcte voorspelling is.

Alle data

	userId	movieId	rating	timestamp
0	182	2571	5.0	1054779786
1	182	2710	4.5	1063284735
2	182	4306	4.0	1054780155
3	195	2571	3.0	974705726
4	195	2710	1.0	974706758
5	195	4306	3.0	994032742
6	204	2571	4.5	1327183462
7	204	2710	5.0	1327185697
8	204	4306	4.0	1327182567
9	376	2571	3.5	1364994024
10	376	2710	1.5	1364994544
11	376	4306	4.0	1364994164
12	542	2571	5.0	1163386800
13	542	2710	0.5	1163387159
14	542	4306	5.0	1163387194

Training set

	userId	movieId	rating	timestamp
1	182	2710	4.5	1063284735
2	182	4306	4.0	1054780155
5	195	4306	3.0	994032742
6	204	2571	4.5	1327183462
7	204	2710	5.0	1327185697
10	376	2710	1.5	1364994544
11	376	4306	4.0	1364994164
13	542	2710	0.5	1163387159
14	542	4306	5.0	1163387194

Test set

	userId	movieId	rating	timestamp
0	182	2571	5.0	1054779786
3	195	2571	3.0	974705726
4	195	2710	1.0	974706758
8	204	4306	4.0	1327182567
9	376	2571	3.5	1364994024
12	542	2571	5.0	1163386800

Testen

- Hoe goed werken voorspelling vanuit de training set op de test set?

Error

$$error = \sum_{i=1}^{n} |prediction_i - score_i|$$

Mean Error

$$ME = \frac{\sum_{i=1}^{n} |prediction_i - score_i|}{n}$$

Mean Squared Error

$$MSE = \frac{\sum_{i=1}^{n} (prediction_i - score_i)^2}{n}$$

Problemen Collaborative Filtering

- Data schaarste
- Verscheidenheid (diversity)
- Lange staart
- Shilling attacks
- Schaalbaarheid
- Grijze en zwarte schapen

Data schaarste

- Cold start

- Lage dichtheid (low density)

Diversity

- Omdat je A leuk vindt, raden we alleen maar A’s aan.

- Moedig exploratie aan

Long Tail

Shilling attacks

- Positieve reviews bij alles van bedrijf A

- Negatieve reviews bij alles van concurrent B

Schaalbaarheid

	User 1	User 2	User 3	...
User 1	1.00	0.20	0.50	
User 2	0.20	1.00	0.21	
User 3	0.50	0.21	1.00	
...				

Grijze en zwarte schapen

- Grijs, er is geen groep waarbij je consistent past.

- Zwart, niemand lijkt op je.

Content-based recommendation

- Komt de inhoud overeen
- Features zijn eigenschappen en niet interacties:
  - Genre
  - Snelheid
  - Prijs
  - Locatie
  - Tijd
  - Taal
  - ...
- Week 5