

Oefentamen Collectieve Intelligentie

...

Vul hier je naam en studentnummer in vóór je begint: 	/ 43 p
---	--------

1. Je mag de vragen in Engels of Nederlands beantwoorden.
2. Dit is een "gesloten boek"-tentamen. Je mag voor het invullen je pen of potlood gebruiken, maar verder niets. Schrijf duidelijk en niet te groot.
3. Leg je studentenkaart (of ander ID met foto) klaar op je tafel. We komen langs om te kijken of je hierboven je naam hebt ingevuld en of deze klopt met je ID.
4. Vul je antwoorden in op dit tentamenblad zelf.
5. Laat het weten als je extra papier of kladpapier nodig hebt.
6. Als je vragen hebt over hoe we iets bedoelen, dan kunnen we dat waarschijnlijk niet beantwoorden zonder een deel van het antwoord weg te geven (maar voel je vrij om het te proberen!).
7. Tip: Het tentamen bevat vragen over veel verschillende onderwerpen. Kijk eerste even het hele tentamen door en begin met de onderwerpen die je het beste beheerst.

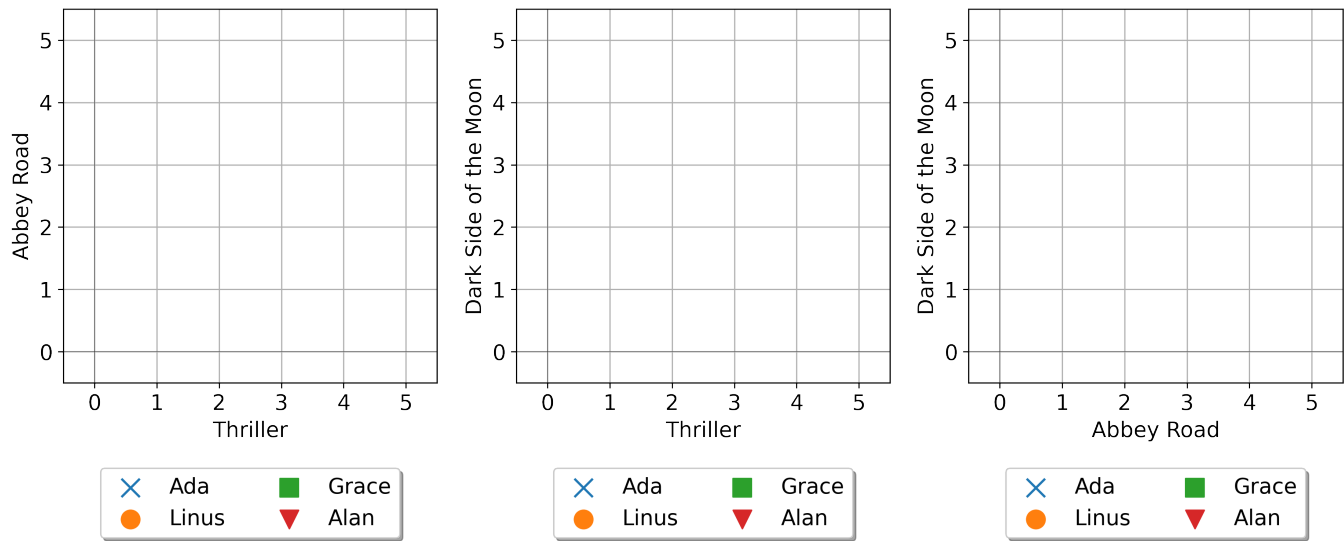
Muziek

Je bent een *collaborative filtering*-algoritme aan het ontwikkelen voor het aanbevelen van muziekalbums. Je hebt de volgende data:

	user	album	rating
0	Ada	Thriller	5.0
1	Linus	Thriller	2.5
2	Grace	Thriller	2.0
3	Alan	Thriller	4.0
4	Ada	Abbey Road	4.0
5	Linus	Abbey Road	2.0
6	Grace	Abbey Road	4.0
7	Alan	Abbey Road	5.0
8	Ada	Dark Side of the Moon	2.0
9	Linus	Dark Side of the Moon	1.0
10	Grace	Dark Side of the Moon	4.0
11	Alan	Dark Side of the Moon	1.0

2 p **Question 1.**

Vul aan de hand van deze data de onderstaande figuren in:



3 p **Question 2.**

Maak voor de bovenstaande data hieronder de utility matrix.

Antwoord

3 p **Question 3.**

Wat is de *manhattan distance* tussen Alan en Linus?

Antwoord

4 p **Question 4.**

We hebben het vermoeden dat van alle gebruikers Ada en Linus het meest op elkaar lijken. Welke similarity-maat reflecteert dit het best (*manhattan similarity*, *euclidean similarity* of *cosine similarity*)? Leg uit waarom. Je hoeft dit niet uit te rekenen, je kan dit beredeneren.

Antwoord

TF-IDF

4 p Question 5.

Ik vroeg ChatGPT om de TF-IDF score voor de term "appel" te berekenen voor drie documenten (A, B en C). ChatGPT kwam met de onderstaande tabel aanzetten. Waarom kan je aan deze tabel meteen zien dat het nooit kan kloppen wat ChatGPT bedacht heeft? Leg uit.

Document	TF (Term Frequency)	IDF (Inverse Document Frequency)
A	0.04	0.00
B	0.02	0.05
C	0.03	0.03

Antwoord

3 p Question 6.

Gegeven de onderstaande twee zinnen, welke heeft een hogere TF-IDF score voor het woord "appel" (ze komen beide uit dezelfde dataset). Leg uit waarom.

1. "De appel valt niet ver van de boom."
2. "De appel viel op het hoofd van Newton omdat hij niet wist dat deze niet ver van de boom zou vallen."

Antwoord

Sagrada Família

3p Question 7.

De volgende data komt van een toeristenwebsite. Het bevat populaire Spaanse steden en keywords die bij de stad passen.

	City	Keywords
0	Barcelona	Gaudí architecture, Beaches, Tapas, Sagrada Família, Cathedral
1	Madrid	Art museums, Plaza Mayor, Tapas, Royal Palace, Flamenco
2	Seville	Moorish architecture, Flamenco, Alcázar, Tapas, Cathedral
3	Valencia	City of Arts and Sciences, Beaches, Paella, Oranges, Modern architecture

Gebruik makend van de *Jaccard index*, welke stad lijkt het meeste op Barcelona (anders dan Barcelona zelf, natuurlijk)? Ter herinnering, de Jaccard index is gedefinieerd als:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Beargumenteer (of reken uit).

Antwoord

Evaluatie

3 p Question 8.

We hebben een recommender system gemaakt voor het aanbevelen van steden aan potentiële toeristen. Om te bepalen of het goed werkt hebben we besloten het systeem online te evalueren aan de hand van historische data. Beschrijf hieronder de stappen voor het online evalueren van een recommender system:

Antwoord

3 p Question 9.

Het evalueren van het recommender system heeft de volgende *confusion matrix* opgeleverd:

	used	unused
recommended	80	20
hidden	120	30

Wat wordt er bedoeld met de termen *used*, *unused*, *recommended* en *hidden*?

Antwoord

3 p Question 10.

Wat is de *precision* gegeven de bovenstaande *confusion matrix*?

Antwoord

2 p **Question 11.**

Voor evaluatie is de precisie van de voorspelling niet de enige relevante maat. Voor de evaluatie kan je ook proberen rekening te houden met andere potentiële doelen van het recommender system. Eén zo'n doel is *serendipity*. Wat wordt hiermee bedoeld?

Antwoord

3 p **Question 12.**

Colaborative filtering werkt over het algemeen niet goed voor het aanbevelen van nieuwe items (een product dat net is toegevoegd aan de catalogus en waar gebruikers nog niet mee te maken hebben gehad). Waarom is dat?

Antwoord

Matrices

3 p Question 13.

$$A = \begin{bmatrix} 2 & 0 & 2 \\ 1 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

We hebben de volgende vergelijking:

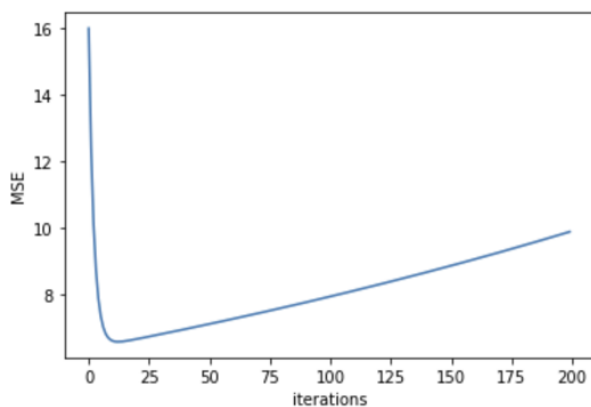
$$A \cdot B = C$$

Wat is de matrix C ?

Antwoord

4 p Question 14.

Bij het runnen van *gradient descent* voor matrixfactorisatie heb je de MSE per iteratie geplot. Eerst gaat de MSE zoals verwacht snel naar beneden, maar daarna gaat deze onverwacht weer omhoog (zie de plot). Waar zou dit door kunnen komen?



Antwoord

Gastcolleges

Je kan drie of vier vragen voor de gastcolleges verwachten...