Tentamen Collectieve Intelligentie

donderdag 29 april 2021 15:00-17:00

1. **Je moet dit tentamen uiterlijk om 17:00 op de website ci.mprog.nl/tentamen inleveren.** (Als je recht op extra tijd hebt, is de deadline 17:30.)
2. De vragen zijn gebaseerd op de gastcolleges, instructievideo's en opdrachten. Naast de informatie op de website mag je ook gebruik maken van alle andere bronnen die je online kan vinden. Geef dan wel aan welke bronnen je hebt gebruikt.
3. Je mag niet samenwerken! Dit is een individueel tentamen.

Doel

Dit tentamen bevat alleen open vragen. Op de meeste vragen is meer dan één antwoord mogelijk. Het gaat er dan ook vooral om hoe je je antwoorden beargumenteerd. Het doel is dat je laat zien dat je de kennis die je hebt opgedaan kan toepassen op een nieuwe situatie.

Je haalt een voldoende voor dit tentamen als je laat zien dat je de relevante informatie uit de colleges en opdrachten weet te gebruiken in je antwoorden, en onwaarheden en irrelevante zijsporen weet te vermijden.

Wat ga je doen?

Je gaat twee praktijkvoorbeelden uitwerken. Voor beide voorbeelden is er een lijst met vragen om te beantwoorden.

Tot slot wordt je gevraagd om de verschillen tussen beide voorbeelden te bespreken.

Praktijkvoorbeeld 1: medicatie[1]

Het volgende praktijkvoorbeeld beschrijft reviewdata voor medicatie. Patiënten met bepaalde aandoeningen hebben hun ervaring gedeeld met de medicatie de ze is voorgeschreven. Deze informatie kan je gebruiken voor het maken van een recommender system dat patiënten bepaalde medicatie aanbeveelt.

Bestudeer het voorbeeld goed. Beantwoord daarna de vragen die eronder staan.

Je kan alle antwoorden geven aan de hand van de data zoals die in dit document wordt beschreven. Je hoeft de data sets niet te downloaden en verder te inspecteren.

Deze dataset bevat de reviews en ratings van patiënten voor bepaalde medicatie. De data set bevat één tabel met :

- drugName (632 unieke waardes): de naam van de medicatie
- condition (1208 unieke waardes): de naam van de aandoening
- review: de review van een patiënt voor de medicatie
- rating: de rating die de patiënt heeft gegeven op een schaal van 1 to 10
- date: de datum van de entry
- usefulCount: het aantal gebruikers dat de review nuttig vond

Voorbeeld tabel (215063 entries):

	uniqueID	drugName	condition	review	rating	date	useful count
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10
4	35696	Buprenorphine	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37
…	…	…	…	…	…	…	…

[1] sources: https://www.kaggle.com/chocozzz/recommendation-medicines-by-using-a-review and https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

35 p **Vraag 1.**

(*Deze vragen gaan over praktijkvoorbeeld 1. Geschatte benodigde tijd voor dit deel: 40 minuten.*)

(a) Hoe zou je een *collaborative filtering* recommender system kunnen maken voor deze data? Geef hierbij genoeg technisch detail:
- Leg voor elke feature in de tabel uit in hoeverre de deze uit de data bruikbaar is voor zo'n systeem.
- Leg uit of bepaalde features nog voorbewerkt moeten worden voordat ze bruikbaar zijn.
- Leg uit welke similarity maat (of maten) je zou gebruiken.
- Leg uit hoe het systeem uiteindelijk aan de hand van voorspelde rating aan een aanbeveling komt.

(b) Hoe zou je een *content-based* recommender system kunnen maken voor deze data? Geef ook hier weer genoeg technisch detail:
- Leg voor elke feauture in de tabel uit in hoeverre de features uit de data bruikbaar zijn voor zo'n systeem.
- Leg uit of bepaalde features nog voorbewerkt moeten worden voordat ze bruikbaar zijn.
- Leg uit welk soort algoritme je gebruikt om tot een aanbeveling te komen.
- Leg uit, indien van toepassing, welke similarity maat (of maten) je zou gebruiken.

(c) Van welke methode verwacht je dat betere voorspellingen, collaborative filtering of content-based filtering?

(d) Zijn er features die gevoelig zouden kunnen zijn voor een *covariate shift*? Beargumenteer

(e) Zo ja, wat zou je kunnen doen om de nadelige effecten van zo'n *covariate shift* te beperken?

(f) Is het theoretische probleem van een *filter bubble* of *fringe bubble* toepasbaar op dit praktijkvoorbeeld? Beargumenteer.

(g) Zo ja, wat zou je kunnen doen om dat te voorkomen?

(h) Is er een risico dat het systeem bevooroordeelde aanbevelingen doet (bijvoorbeeld op grond van geslacht of etniciteit)? Beargumenteer

(i) Zo ja, wat zou je kunnen doen om dat te voorkomen?

Praktijkvoorbeeld 2: Uitzendbureau[2]

Het volgende voorbeeld bevat informatie van een uitzendbureau. Hiermee zou een recommender system voor het koppelen van werkzoekenden aan banen gemaakt kunnen worden.

Bestudeer het voorbeeld goed. Beantwoord daarna de vragen die eronder staan.

Je kan alle antwoorden geven aan de hand van de data zoals die in dit document wordt beschreven. Je hoeft de data sets niet te downloaden en verder te inspecteren.

De data bevat twee tabellen. De eerste tabel bevat alle vacatures bekend bij het uitzendbureau (zowel de huidige vacatures als die uit het verleden). De tabel bevat:

- status: staat de vacature op dit moment open?
- title: titel van de vacature
- position: functienaam
- company: de naam van het bedrijf
- city: de gemeente waar de werknemer zal gaan werken
- job description: een uitgebreide beschrijving van de aanstelling (hiervan wordt in het voorbeeld hieronder slechts een kort fragment getoond)
- salary: loon
- listing start: begindatum aanstelling
- listing end: einddatum aanstelling
- employment type: aanstellingstype (deeltijd, voltijd, etc.)
- education required: vereist opleidingsniveau

De tweede tabel bevat de volledige werkgeschiedenis van alle werkzoekende in de database van het uitzendbureau. Deze bevat:

- applicant id: de unique id van de werknemer
- position name: de functienaam behorende bij een eerdere aanstelling
- employer name: de naam van de toemalige werkgever
- city: de gemeent van de aanstelling
- start date: de startdatum van de aanstelling
- end date: de einddatum van de aanstelling
- job description: een uitgebreide beschrijving van de aanstelling (hiervan wordt in het voorbeeld hieronder slechts een kort fragment getoond)
- salary: het verdiende inkomen

Voorbeeldtabellen op de volgende pagina.

[2] source: https://www.kaggle.com/c/job-recommendation/

Voorbeeld tabel 1 (*jobs*, 84083 rijen):

Job ID	Status	Title	Position	Company	City	Job Description	Salary	Listing Start	Listing End	Employment Type	Education Required
134273	open	Assistant Store Manager @ King's Food Markets	Assistant Store Manager	King's Food Markets	Mendham	[...] Assistant Store Manager [...]	45000	12-05-2014	01-04-2015	Part-Time	Not Specified
134274	open	Store Associate - Retail Sales (Customer Service) @ ALDI	Store Associate - Retail Sales (Customer Service)	ALDI	Onalaska	Store Associate [...]	NA	12-05-2014	01-04-2015	Full-Time/Part-Time	High School Diploma
134275	open	Macy's Seasonal Retail Commission Sales Women's Shoes, Part Time - Skokie, IL - Old Orchard Mall @ Macy's	Macy's Seasonal Retail Commission Sales Women's Shoes, Part Time - Skokie, IL - Old Orchard Mall	Macy's	Skokie	Sales Associate is responsible for [...]	NA	12-05-2014	01-04-2015	Seasonal	Not Specified
134276	open	Part Time Clerks Needed @ Kroger - Louisville	Part Time Clerks Needed	Kroger - Louisville	Louisville	Part Time Store Clerks Needed - All Departments [...]	25000-35000	12-05-2014	01-04-2015	Part-Time	Not Specified
134277	open	Flex (Part Time) Merchandiser - Mullins, SC @ Coca-Cola Bottling Company Consolidated	Flex (Part Time) Merchandiser - Mullins, SC	Coca-Cola Bottling Company Consolidated	Mullins	[...] responsible for stocking and [...]	NA	12-05-2014	01-04-2015	Part-Time	Not Specified
134278	open	Child Psychotherapist @ CATCH, Inc.	Child Psychotherapist	CATCH, Inc.	Philadelphia	[...] intake assessments for children. Individual therapy and [...]	66000-75000	12-05-2014	01-04-2015	Full-Time/Part-Time	Master's Degree
134281	open	HR Administrator @ New York Community Bank	HR Administrator	New York Community Bank	Elizabeth	Provides administrative support for the Employee Relations [...]	NA	12-05-2014	01-04-2015	Part-Time	Not Specified
…	…	…	…	…	…	…	…	…	…	…	…

Voorbeeld tabel 2 (*applicant job history*, 18642 rijen):

Applicant.ID	Position.Name	Employer.Name	City	Start.Date	End.Date	Job.Description	Salary
10007	Bartender	15 Romolo	San Francisco	2010-02-01	2011-04-30	Serve luxury cocktails and wines in an intimate [...]	NA
10008	non-food stocker, photo	Kroger	Cincinnati			stocked non food grocery items, kept aisle neat, did [...]	6.25
10008	Registrar	Mercy Health Physicians	Cincinnati	2012-10-01		incoming call, registration, problem solve, office, clerical [...]	13.53
10008	Registrar	The Christ Hospital	Cincinnati	2006-10-01	2012-08-27	registration, EMR, calls, referral, clerical/office work	12.5
10008	cashier/baker	The Cheesecake Factory	Cincinnati	2004-11-01	2006-09-01	cashier, all take out call in orders, prepared desserts/beverages [...]	10
10008	cashier/asst.manager	Penn station	Cincinnati	2000-06-01	2004-10-01	opened/closed store, prepped food, cook, cashier, supervisor	9.5
1001	Sales Associate	Athleta	Santa Monica	2013-07-01	2013-01-01	Assist customers on sales floor and in fitting rooms. Provide a [...]	8.75
…	…	…	…	…	…	…	…

35 p **Vraag 2.**

(*Deze vragen gaan over praktijkvoorbeeld 2. Geschatte benodigde tijd voor dit deel: 40 minuten.*)

(a) Hoe zou je een *collaborative filtering* recommender system kunnen maken voor deze data? Geef hierbij genoeg technisch detail:
- Leg voor elke feature in beide tabellen uit in hoeverre de deze uit de data bruikbaar is voor zo'n systeem.
- Leg uit of bepaalde features nog voorbewerkt moeten worden voordat ze bruikbaar zijn.
- Leg uit welke similarity maat (of maten) je zou gebruiken.
- Leg uit hoe het systeem uiteindelijk aan de hand van voorspelde rating aan een aanbeveling komt.

(b) Hoe zou je een *content-based* recommender system kunnen maken voor deze data? Geef ook hier weer genoeg technisch detail:
- Leg voor elke feauture in beide tabellen uit in hoeverre de features uit de data bruikbaar zijn voor zo'n systeem.
- Leg uit of bepaalde features nog voorbewerkt moeten worden voordat ze bruikbaar zijn.
- Leg uit welk soort algoritme je gebruikt om tot een aanbeveling te komen.
- Leg uit, indien van toepassing, welke similarity maat (of maten) je zou gebruiken.

(c) Van welke methode verwacht je dat betere voorspellingen, collaborative filtering of content-based filtering?

(d) Zijn er features die gevoelig zouden kunnen zijn voor een *covariate shift*? Beargumenteer

(e) Zo ja, wat zou je kunnen doen om de nadelige effecten van zo'n *covariate shift* te beperken?

(f) Is het theoretische probleem van een *filter bubble* of *fringe bubble* toepasbaar op dit praktijkvoorbeeld? Beargumenteer.

(g) Zo ja, wat zou je kunnen doen om dat te voorkomen?

(h) Is er een risico dat het systeem bevooroordeelde aanbevelingen doet (bijvoorbeeld op grond van geslacht of etniciteit)? Beargumenteer

(i) Zo ja, wat zou je kunnen doen om dat te voorkomen?

20 p **Vraag 3.**

(*Geschatte benodigde tijd voor dit deel: 30 minuten.*)

Als het goed is zitten er verschillen tussen de antwoorden voor beide praktijkvoorbeelden.

Benoem de vier belangrijkste verschillen tussen beide recommender systems zoals je ze hierboven hebt beschreven. Duidt welke aspecten van het praktijkvoorbeeld voor deze verschillen zorgen.