Tentamen Collectieve Intelligentie

donderdag 30 mei 2020 11:00-17:00

1. Dit tentamen bestaat uit een aantal open vragen. In veel gevallen zijn er meerdere antwoorden mogelijk. De redenering is dan belangrijker dan het antwoord.
2. Bij elke vraag staat een maximum aantal woorden vermeld. Ga hier niet overheen. De grens is heel ruim genomen dus het antwoord kan meestal een stuk beknopter.
3. Je moet dit tentamen uiterlijk om 17:00 op de website ci.mprog.nl/tentamen inleveren.
4. Je hebt ruim de tijd voor dit tentamen. Neem dus ook af en toe pauze.
5. De vragen zijn gebaseerd op de gastcolleges, intructievideo'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.
6. Je mag niet samenwerken! Dit is een individueel tentamen.
7. De laatste vraag (vraag 5) is aanzienlijk langer dan de andere vragen. Hou hier rekening mee met je planning.

10 p **Question 1.**

(a) Personalisatie is gedefinieerd als "Adapting to user preferences". Waarom zou je dat willen? Geef minstens één voorbeeld van een toepassing van personalisatie die geen commercieel oogmerk heeft.
[max 50 woorden]

(b) Noem een voorbeeld van expliciete feedback en een voorbeeld van impliciete feedback.
[max 50 woorden]

(c) Wanneer het 'Cold Start'-probleem optreedt kan een recommender system alleen algemene suggesties doen. Naarmate een recommender system dan meer over de gebruikers weet worden geleidelijk specifiekere suggesties gedaan. Geef een voorbeeld van een toepassingsgebied waar deze strategie niet werkt en beargumenteer welke strategie dan wel wordt gekozen.
[max 50 woorden]

20 p **Question 2.**

(a) Voordat je een recommender system kunt evalueren moet eerst duidelijk zijn wat je met dit systeem probeert te bereiken. Geef minstens drie verschillende voorbeelden van doelen die een recommender system kan hebben en geef bij ieder doel een voorbeeld van een bijbehorend recommender system.
[max 100 woorden]

(b) Leg uit welke drie manieren er zijn om een recommender system te testen en benoem welke voor- en nadelen ieder van deze methoden heeft.
[max 100 woorden]

(c) Leg uit waarom je je data zou onderverdelen in een training- en een testset en geef één eigenschap waaraan de twee subsets van je data moeten voldoen om valide uitspraken te kunnen doen.
[max 100 woorden]

(d) Leg uit wat een covariate shift is en noem een voorbeeld. Leg aan de hand van dit voorbeeld uit wat je moet doen met het model dat ten grondslag ligt aan je recommender system als je een covariate shift detecteert?
[max 50 woorden]

10 p **Question 3.**

(a) Hoe is het probleem van filter bubbles gerelateerd aan collaborative filtering?
[max 100 woorden]

(b) Zijn filter bubbles een echt probleem? Waarom (niet)? Noem tenminste één argument uit de podcast van Judith Möller.
[max 100 woorden]

(c) Filter bubbles zijn niet in alle landen een even groot probleem. Noem minstens één concreet verschil tussen Nederland en de Verenigde Staten.
[max 100 woorden]

20 p **Question 4.**

Voor een muziekschool willen we leerlingen werven door een recommender system suggesties te laten geven voor welk muziekinstrument het beste bij potentiële leerlingen past. Om zo'n systeem te kunnen bouwen worden alle huidige leerlingen geïnterviewd. Hierbij noteert men met welke muziekinstrumenten ze ervaring hebben en wat ze daar van vonden.

Geef aan hoe je dit systeem zou bouwen zodat het gebruik maakt van collaborative filtering en geef ook aan hoe je dit aan kan pakken op een content-based filtering manier, of beargumenteer waarom recommender systems hier minder effectief zijn.

[max 200 woorden]

Vraag 5 op de volgende pagina!

30 p **Question 5.**

In de toekomst ga je een recommender system bouwen voor de website Yelp. Het doel is om bedrijven aan te raden waar mensen eventueel in geïnteresseerd kunnen zijn. Als je bijvoorbeeld naar restaurants zoekt krijg je een lijstje met 30 restaurants te zien die jij interessant zou kunnen vinden:

https://www.yelp.nl/search?cflt=restaurants

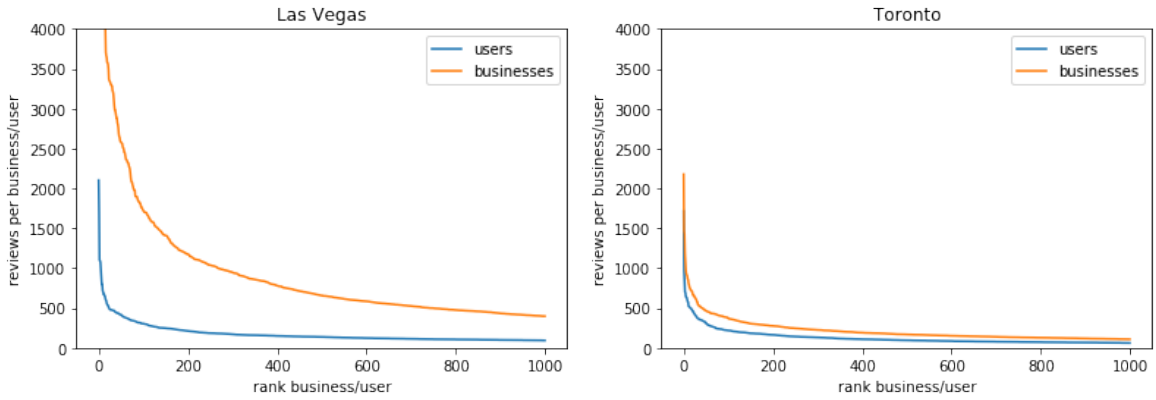
Nu gaat dat vaak aan de hand van locatie, maar de bedoeling is dat meer te personaliseren.

Om een idee te krijgen hoe we dat kunnen doen, gaan we eerst eens naar de data kijken. De data van Yelp staat hier beschreven:

https://www.yelp.com/dataset/documentation/main

Deze dataset bevat reviews voor bedrijven. Bovendien bevat de data ook veel inhoudelijke informatie over de bedrijven zelf.

Laten we eerst eens naar de reviews kijken. De onderstaande plots laten, voor de steden Toronto en Las Vegas, de hoeveelheid reviews per gebruiker en bedrijf zien, gesorteerd van hoog naar laag.



(a) Gegeven deze plots, denk je dat user based of item based collaborative filtering beter werkt? Beargumenteer.
[max 100 woorden]

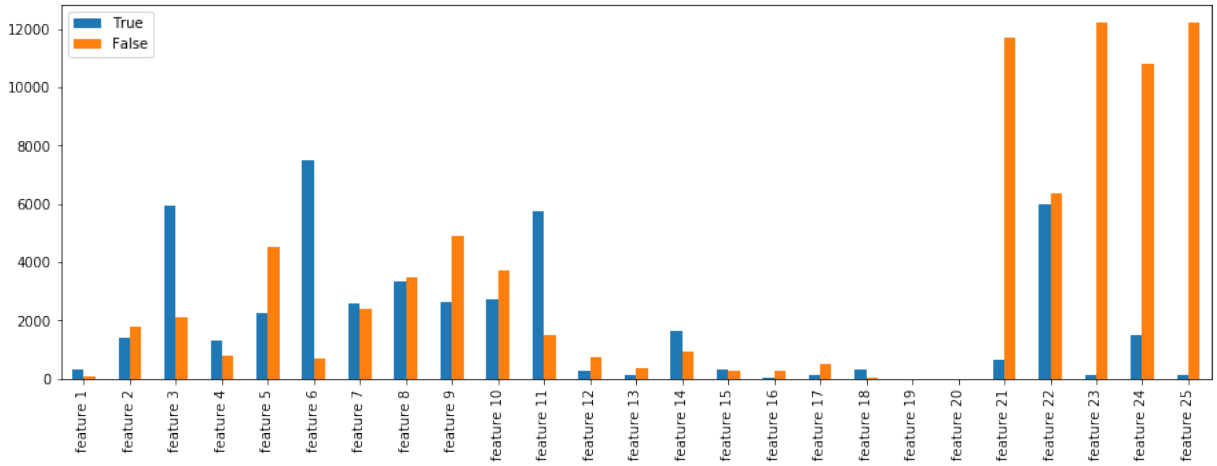
(b) Het antwoord op bovenstaande vraag is natuurlijk op z'n best een indicatie. Als je zeker zou willen weten welke vorm van filtering beter werkt, hoe zou je dat moeten testen?
[max 50 woorden]

(c) Verwacht je dat het verschil tussen user en item based filtering voor beide steden hetzelfde is? Zo nee, voor welke verwacht je dat het verschil het grootste is? Waarom?
[max 50 woorden]

(d) Voor beide steden geldt dat je waarschijnlijk (ook) content based filtering zou willen gebruiken. Waarom is dat? Geef tenminste één reden.
[max 100 woorden]

(e) De data bevat onder andere een hoop binaire features (vooral onderdeel van 'attributes'). Bijvoorbeeld: is

er een parkeergarage (True/False), of (als het om restaurant gaat) wordt er alcohol geschonken (True/False)? Hieronder hebben we, voor een deel van de data, de distributie van een aantal van deze binaire features geplot. We hebben voor deze vraag de echte labels expres achterwege gelaten en ze feature 1 t/m 25 genoemd. (Dit is trouwens een versimpelde voorstelling: de data bevat nog veel meer features dan dit. Bovendien wisselen de features per bedrijfssoort.)



Je kan in principe een combinatie van al deze features gebruiken voor content based filtering. Maar, om het systeem efficiënt te houden zouden we ons willen beperken tot een deel van de features. Gegeven de bovenstaande plot, welke features (geef het nummer) kan je het beste achterwegen laten? Waarom? [max 100 woorden]

(f) Als je een content based algoritme zou gaan implementeren is het altijd goed om eenvoudig te beginnen. Kijk nog even goed naar de beschrijving van de Yelp data:
https://www.yelp.com/dataset/documentation/main
Hierin staan een aantal features bij naam genoemd. Welke van deze features lijken je het meest geschikt voor een eerste prototype? Beargumenteer waarom je je juist op die features zou richten.
[max 100 woorden]

(g) Welke similiarity-maten zou je gebruiken voor het vergelijken van deze features? Waarom?
[max 50 woorden]