Deelverslag 1 Groepsproject Collectieve Intelligentie

Groepsleden:

- Rabie Afqir (**11606843**)
- Björn Out (**12567930**)
- Frank Tamer (**12248738**)
- Gino Pennasilico (**12393819**)

1.0 Taakverdeling

Deze week willen wij het prototype af hebben zodat wij volgende week kunnen evalueren wat de uitkomsten van de algoritmes zijn. Rabie en Björn zullen aan het SVM-algoritme (content-based) werken en Frank en Gino zullen het item based collaborative filtering algoritme verzorgen. De taken (documenteren, research doen en het programmeren) zullen per deelgroep geregeld worden. We kennen elkaar al goed dus de onderlinge coördinatie en afstemming zal goed komen.

1.1 Structuur

A. Hoe ziet de review data er uit? Welke features? [max 100 woorden]
De review data bevat de volgende features:

	business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	dJ0R-XT78LUQeNHQkD-G9g	0	2018-08-14 04:03:02	0	qb2EVdmVNvw3D0kBMN6Xrg	5.0	Best place to get ice cream. They have only tw...	0	hXydWH25S92Hjl5hmWRSyA
1	Q_0eGl-aElqHKukHvmLdwA	0	2015-12-05 23:21:39	0	Ecr_pKR7786kmcLVXLp5NA	1.0	Sorry to say, Nelias did not live up to the ot...	0	vo6vLeHoPl_h-Vt-YHs9_A
2	Q_0eGl-aElqHKukHvmLdwA	0	2015-08-07 19:33:13	0	Ro6-JL0KCS5JULXUNRST-w	1.0	Wouldn't give it one star if I could. This pla...	0	i1qyYL4fpAel8Ljt4WaZ3g

Er is te zien dat elke review gekoppeld is aan 1 gebruiker (de gebruiker die een rating geeft) en aan 1 business (de business die beoordeeld wordt). De manier van beoordelen door gebruikers wordt gedaan aan de hand van het geven van sterren. Waarbij 1 ster de laagst mogelijke en 5 sterren de hoogste beoordeling is. Tot slot bestaat al het overige van review data uit features die gedetailleerde informatie weergeeft over een beoordeling. Bijvoorbeeld de datum, tekst en scores over hoe grappig, cool of nuttig een beoordeling is.

B. Hoe ziet de user data er uit? Welke features? [max 100 woorden]
De user data bestaat uit een aantal features die algemene informatie bevatten over een gebruiker. Je kan hierbij denken aan de user_id, de naam van de gebruiker, hoe vaak de

gebruiker een beoordeling achter heeft gelaten en het gemiddelde aantal sterren gegeven door de gebruiker. Verder geven de andere features inzicht in de interactie tussen de gebruiker en andere gebruikers op het platform. Denk hierbij aan hoeveel en wat voor soort complimenten de gebruiker heeft ontvangen en gegeven aan anderen.

Alle features van de user data:

```
['average_stars', 'compliment_cool', 'compliment_cute',
 'compliment_funny', 'compliment_hot', 'compliment_list',
 'compliment_more', 'compliment_note', 'compliment_photos',
 'compliment_plain', 'compliment_profile', 'compliment_writer', 'cool',
 'elite', 'fans', 'friends', 'funny', 'name', 'review_count', 'useful',
 'user_id', 'yelping_since'],
```

C. Hoe ziet de business data er uit? Welke features? Zijn er features die hetzelfde zijn voor alle businesses? Zijn er features die verschillen per business? [max 100 woorden]

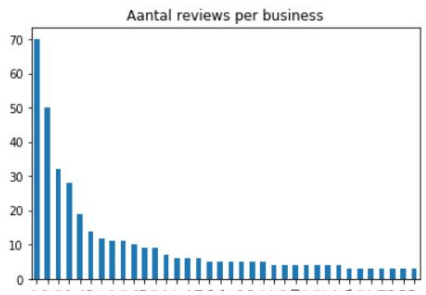
De business data geeft informatie weer met betrekking voor de businesses. Zo is er algemene informatie beschikbaar zoals het adres, de naam van de business de openingstijdens en in welke voor categoriën een business hoort. Daarnaast zijn er ook features die het aantal beoordelingen en gemiddelde aantal sterren van een business heeft. Deze zouden nuttige input kunnen zijn voor de algoritmes. Sommige features zoals ‘attributes’ en ‘categories’ verschillen per business terwijl features zoals state of city hetzelfde zullen zijn voor businesses in dezelfde stad / staat.

Alle features van de business data:

```
['address', 'attributes', 'business_id', 'categories', 'city', 'hours',
 'is_open', 'latitude', 'longitude', 'name', 'postal_code',
 'review_count', 'stars', 'state'],
```

1.2 Plots

Deze plots zijn gemaakt op basis van gegevens van één stad, omdat dit veel rekenkracht scheelt ten opzichte van het plotten van de hele data.

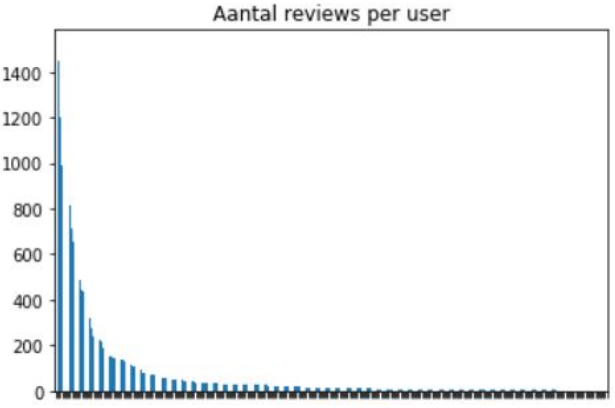


A. Plot reviews en leg verdeling uit:

Hier in de figuur te zien, zijn het aantal reviews dat bedrijven in Ambridge hebben gekregen. Er is te zien dat deze een ‘tailed distribution’ heeft, kenmerkend met zijn long tail vorm. Er is dus een handjevol bedrijven dat daadwerkelijk veel reviews heeft mogen ontvangen, en een groot deel van de stad moet het doen met maar 10 reviews. Dit omdat populairdere bedrijven natuurlijk vaker bekeken worden en zo nog meer bezocht worden en een grotere kans op een rating hebben.

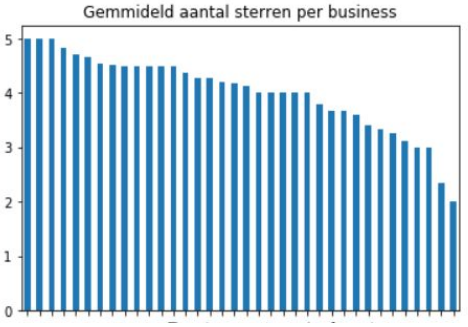
B. Plot user data en leg verdeling uit:

Hier in de figuur zijn het aantal reviews gegeven per user, gegeven aan bedrijven in Ambridge.Ook deze plot heeft een ‘heavy tailed distribution’ Veel gebruikers hebben weinig of maar één review gegeven binnen Ambridge. Dit komt doordat gebruikers van nature niet altijd een review achterlaten, en ook omdat er veel accounts amper gebruikt worden. Het komt weleens voor dat iemand een account maakt om simpelweg één review te geven, omdat bijvoorbeeld de ervaring erg goed was. Sommige gebruikers reviewen daarentegen erg veel. Mogelijk hoort dit bij hun werk of vinden ze dit een erg leuke bezigheid.



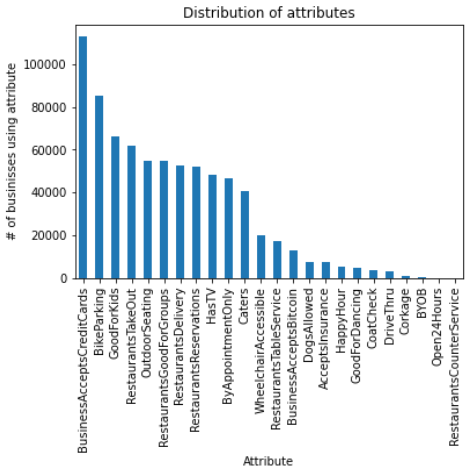
C. Plot business data en leg verdeling uit:

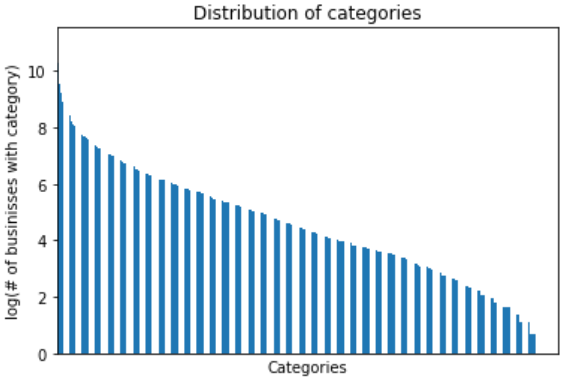
Hier rechts in de grafiek is de verdeling te zien van het gemiddeld aantal sterren dat een bedrijf in Ambridge heeft mogen ontvangen. Er is een vrij neerwaartse trend te zien, waarbij het gemiddelde vooral tussen de 2 en de 5 sterren blijft. Enkele bedrijven hebben 5 sterren, wellicht komt dit door een laag aantal reviews. Dat de grens van 2 sterren hier niet overschreden wordt, valt wellicht te wijten aan het feit dat er meerdere reviews gegeven worden, die het gemiddelde hoger dan 2 houden.



De staafgrafiek rechts toont per *attribute* hoe veel bedrijven dit *attribute* ingevuld hebben. Te zien is dat slechts de eerste drie echt veel ingevuld zijn. Vanaf de *attribute* ‘RestaurantsTakeOut’ is er onder andere sprake van *attributes* die specifiek voor bepaalde typen bedrijven bedoeld zijn.

De laatste plot hieronder geeft de distributie van de verschillende bedrijfscategorieën weer. Dit is op een logaritmische schaal geplot, omdat er anders niets te zien is. Te zien is dat er enkele categorieën zijn die zéér vaak gebruikt zijn, terwijl sommigen bijna niet gebruikt worden.





1.3 Verwachtingen

A. Verwacht je dat Collaborative Filtering (CF) goed werkt? Waarom wel/niet? Is user based of item based beter? [max 150 woorden]

Wij verwachten dat Collaborative Filtering niet heel goed zal werken, omdat de data erg *sparse* is: gebruikers geven aan slechts enkele restaurants een review en restaurants zullen ook van slechts een klein deel van de gebruikers een review krijgen. De overlap tussen reviewers van één bedrijf ten opzichte van de reviewers van een ander bedrijf en de overlap van gereviewde bedrijven van één gebruiker ten opzichte van de gereviewde bedrijven van een andere gebruiker zal hierdoor erg klein zijn. Als gevolg daarvan zullen bedrijven en gebruikers onderling slecht met elkaar te vergelijken zijn op basis van reviews. Wel zal *item based CF* waarschijnlijk beter werken dan *user based CF*, omdat er minder bedrijven zijn dan gebruikers en bedrijven gemiddeld gezien meer reviews zullen ontvangen dan een gebruiker gemiddeld gezien zal geven. Er is hierdoor meer kans dat features (reviewers) tussen verschillende items (bedrijven) zullen overlappen.

B. Zo niet, is er een subset van de data waarvoor je wel verwacht dat CF werkt? [max 150 woorden]

Ja. Op het moment dat je alleen de meest populaire bedrijven, dat wil zeggen de bedrijven met de meeste reviews, meeneemt, zal CF aanzienlijk beter werken. Bedrijven met veel reviews hebben onderling veel meer kans om overlappende features (gebruikers die het bedrijf hebben gereviewd) te hebben. Hierdoor is een groot deel van het probleem geschetst bij A. opgelost, en zal het systeem aanzienlijk betere suggesties doen. Uiteraard sluit je hiermee wel kleinere en minder populaire bedrijven uit, die wel goed zouden kunnen zijn. Deze uitsluiting is absoluut ongewenst (één van de doelen van Yelp is gebruikers nieuwe en niche dingen laten ontdekken). Wat dat betreft is deze aanpak geen oplossing voor het probleem, maar qua harde metrics zou het algoritme beter presteren.

C. Verwacht je dat Content-based Filtering (CbF) goed werkt? Waarom wel/niet? [max 150 woorden]

Wij verwachten dat CbF wel goed zal werken, voornamelijk omdat bedrijven veel overlappende features hebben, en omdat deze features de bedrijven ook goed omschrijven. De verschillende typen bedrijven hebben ook onderling weer verschillende extra attributen, wat het ook goed mogelijk maakt om bedrijven van hetzelfde type met elkaar te vergelijken. Doordat bedrijven op basis van Content Based filtering allemaal goed met elkaar te vergelijken zullen zijn, kan er zelfs op basis van slechts één review van een gebruiker al gepersonaliseerd worden, en is het probleem met *sparse data* zoals omschreven bij B. opgelost.

D. Welke features denk je dat nuttig kunnen zijn voor CbF? Waarom? [max 150 woorden]

De verwachting is dat de features *categories*, *attributes*, *stars*, en *city* erg nuttig zullen zijn in ons CbF-algoritme. Aan de hand van de *city* kunnen we gebruikers alleen bedrijven

aanraden in de buurt van waar zij wonen of vaak zijn. *Categories* maakt het mogelijk om onderscheid tussen verschillende typen bedrijven, waar gebruikers mogelijk wel of niet in geïnteresseerd zijn. *Attributes* maken het mogelijk bedrijven met dezelfde categorieën onderling met elkaar te vergelijken en de aanbevelingen per persoon specifieker te maken (bijvoorbeeld: de ene gebruiker vindt parkeergelegenheid heel belangrijk, de ander niet). De *stars* rating maakt het mogelijk om een indicatie van de kwaliteit van dienstverlening van het bedrijf mee te nemen bij de aanbevelingen, waarbij we ons uiteraard richten op het aanbevelen van de betere bedrijven.

1.4 Plannen

A. Welke algoritmes (minstens 2) wil je testen? [max 75 woorden per algoritme (of 150, als je een algoritme wil gebruiken dat niet in de cursus is behandeld en dus uitgelegd moet worden)]

Wij willen als content-based filtering algoritme Support Vector Machines (SVM's) modelleren. We volgen momenteel allemaal het vak Toegepaste Machine learning dus we zijn hier allemaal bekend mee. We zullen een lineair SVM gebruiken om zo te laten zien welke wel en niet aangeraden horen te worden. Support vector machines zijn een techniek uit de machine learning, en proberen data op te delen in classificaties door een zo groot mogelijke 'buffer' tussen twee klassen te zoeken. Vervolgens wordt nieuwe data afhankelijk van aan welke kant van de buffer de data ligt geclassificeerd. In dit geval zijn de klassen 'wel aanraden' en 'niet aanraden', waarbij verschillende thresholds qua rating worden gebruikt voor het wel danwel niet aanraden. Om verschillende soorten bedrijven aan te raden, zullen we de aan te raden bedrijven selecteren op basis van interne gemiddelde Jaccard similarity.

Daarnaast zullen we gebruik maken van item-based collaborative filtering. Hier kiezen we voor omdat we verwachten dat er sprake is van een long tail in de review data. Hierdoor is het beter om item-based collaborative filtering te werken dan user-based. Een bedrijf zal over het algemeen meer informatie bevatten dan een user.

Verder zullen we een Cold-Start approach uitwerken op basis van minimum confidence interval, met inachtname van een aantal randvoorwaarden, zoals diversiteit in aanbevelingen. Deze aanpak zou ingezet kunnen worden op het moment dat de andere twee aanpakken niet mogelijk zijn of niet goed zullen werken, bijvoorbeeld wanneer een user nog maar één review heeft gegeven.

B. Wat zijn technische aspecten om rekening mee te houden? Zijn er features die niet altijd aanwezig zijn? Ga je NLP technieken gebruiken die niet altijd betrouwbaar zijn? [max 150 woorden]

Een aantal features zijn soms niet aanwezig. Denk hierbij aan de openingstijden en de attributen van een plaats. Hierdoor is het soms lastig om te weten wat we iemand kunnen aanraden. Aan de andere kant kunnen we wel op een makkelijke manier dit oplossen door van de attributen alleen de booleans te encoden en te gebruiken. Ook kunnen we eventueel bepaalde bedrijven minder snel aanraden wanneer er weinig over de zaak bekend is.

C. Hoe ga je testen of het werkt? Welke evaluatietechnieken ga je gebruiken? Hoe sluiten die aan bij de doelen van Yelp? Zijn er dingen die je zou willen testen maar waarbij het testen om technische/praktische redenen niet mogelijk zijn? [max 150 woorden]

De evaluatiemethode voor het item-based collaborative aanbevelen zal worden gedaan door te kijken naar een expected rating en daar dan de RMSE en de MAE van te berekenen. Dit sluit aan bij de doelen van Yelp door een zo goed mogelijke aanbeveling te doen aan de gebruiker. Iets wat we helaas niet kunnen doen is het voorspellen van een rating d.m.v. een SVM. Een SVM is namelijk een classifier waardoor we niet direct numerieke voorspellingen kunnen doen. Wel zullen we kijken naar de precision-recall curves in vergelijking tot enkele nader te bepalen baselines. Dit zullen wij ook bij het item-based collaborative algoritme doen. Bij beide algoritmen zullen we de thresholds voor het doen van een aanbeveling variëren, om zo tot de best mogelijke oplossing te komen. Ten slotte zullen we de Jaccard similarity vergelijken met een baseline, om een idee te krijgen van de diversiteit in de aangeraden bedrijven.