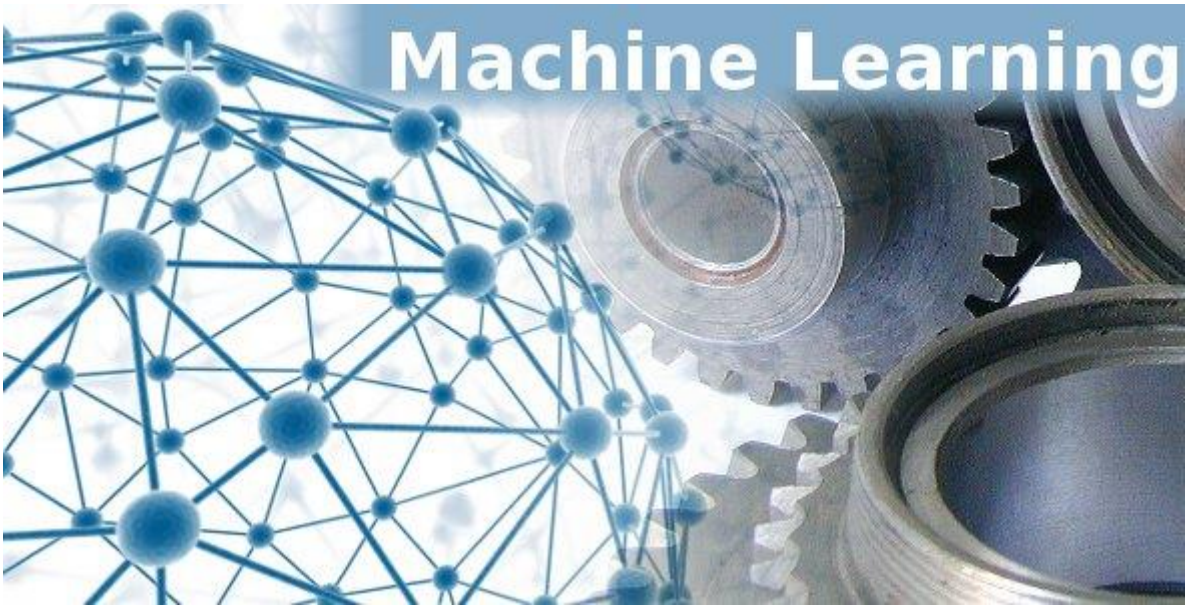
Machine learning

Opdracht 3 ANALYZING AND VISUALIZING BIG DATA



Namen:	Vincent Beltman & Mike Holtkamp
Begeleiders:	Jan Stroet & Evert Duipmans
Klas	EIN3vBDa
Datum	28-6-2015

Inhoudsopgave

1 De opdracht

Deze opdrachten worden uitgevoerd op een set met daarin films, gebruikers en beoordelingen. Het eerste gedeelte van de opdracht is het bepalen van de top 5 filmgenres voor een aantal categorieën zoals de leeftijd en het beroep. In het tweede gedeelte van de opdracht staat het clusteren en het doen van aanbevelingen centraal.

2 Dataset

De verkregen dataset bestaat uit vier verschillende bestanden.

- Het Genre-bestand. Hierin staan de genres gekoppeld aan hun ID.
- Het User-bestand. Hierin staat de informatie over alle users.
- Het Item-bestand. Hierin staat de informatie over alle movies.
- Het Rating-bestand. Hierin staan alle ratings, de user_id's de movie_id's

2.1 Genre

Naam	Uitleg
Genre	De naam van het genre bijvoorbeeld Comedy of Drama.
Genre-Index	Deze index wordt in het Item-bestand gebruikt om te bepalen bij welke genres de film hoort.

2.2 User

Naam	Uitleg
User_id	Het unieke ID van de gebruiker. Dit ID wordt gebruikt bij de rating om te weten welke gebruiker het is.
Leeftijd	De leeftijd van de gebruiker deze wordt gebruikt voor een van de top 5 bepalingen. Deze wordt dan ook omgezet naar een range van leeftijden. 24 jaar wordt bijvoorbeeld 20-30.
Beroep	Het beroep van de gebruiker. Deze wordt gebruikt bij een van de top 5 bepalingen.
Postcode	De postcode van het woonadres van de gebruiker. Deze wordt gebruikt voor het bepalen van een top 5.

2.3 Item / Movie

Naam	Uitleg
Film_id	Het unieke ID van de film om de film te kunnen identificeren.
Title	De titel van de film bijvoorbeeld Toy story.
Release date	De datum wanneer de film is uitgegeven bijvoorbeeld 1 Jan 1995.
Video release date	Zie bovenstaande maar dan voor video.
IMDB URL	De URL naar de online film database IMDB. Deze kan eventueel gebruikt worden om extra informatie op te halen.
Genres	Er zijn 19 waardes wanneer een waarde op een index een 1 is valt de film in dat genre. Een film kan in meerdere genres voorkomen. Index 5 in dit bestand is gelijk aan het genre 0, 6 is gelijk aan 1 enz.

2.4 Ratings

Naam	Uitleg
User_id	Het ID van de user die de film een rating gegeven heeft.
Movie_id	Het ID van de film die door de user is gerate.
Rating	De beoordeling die de gebruiker die film gegeven heeft. Het bereik van het cijfer is van 0 t/m 5. Het zijn hele getallen.
Tijdstip	Een timestamp van wanneer de gebruiker de rating geplaatst heeft.

3 Libraries

In deze opdracht wordt er gebruik gemaakt van twee externe libraries buiten sklearn en numpy om. De eerste library is pprint. Deze library wordt gebruikt om grote dictionaries overzichtelijk uit te printen. Als deze library niet aanwezig is, wordt alles op de normale manier en onoverzichtelijk uitgeprint. Deze library kan worden geïnstalleerd met “pip install pprint” of “easy_install pprint”. Voor meer informatie gaat naar https://docs.python.org/2/library/pprint.html.

De tweede library is scipy. Deze library wordt gebruikt voor de pearson correlation. Als deze library niet aanwezig is, wordt onze eigen implementatie van de pearson correlation gebruikt. Echter is het aangeraden om die van scipy te gebruiken in plaats van die van ons. Deze is net iets sneller. Scipy kan worden geïnstalleerd door middel van “pip install pprint” of “easy_install pprint”. Voor meer informatie ga naar https://pypi.python.org/pypi/scipy.

4 Deel 1 van de opdracht (Analyse van de dataset)

Bij het eerste gedeelte van de opdracht stond het maken van een top 5 centraal. Om de gevraagde onderdelen te kunnen maken hadden we de volgende informatie nodig:

- Genres
- Users
- Movies
- Ratings

Daarnaast zouden we de beroepen nog kunnen inlezen maar we hebber er voor gekozen om dit niet te doen omdat deze informatie ook staat in het User-bestand. Omdat we gebruikers en films op basis van hun ID eenvoudig moeten kunnen vinden, hebben we er voor gekozen om een dictionary te gebruiken met het ID als de key. De eerste stap om tot het resultaat te komen is het inlezen van de hierboven genoemde dataset. Daarna lopen we door de ratings heen en zetten voor de verschillende top 5 items bijvoorbeeld

Geslacht
M {
    Comedy: 15,
    Drama:10
}
F{
    Romance: 999,

    Action: 1
}
Leeftijd

[0 - 10] {
    Children: 80,
    Animation: 10,
    Fantasy: 20

}

We hebben er voor gekozen om maar 1 keer door de dataset heen te lopen en voor elke top 5 criteria gelijk het aantal te zetten omdat ons dit de meest efficiënte manier leek. Vervolgens worden de lijsten aangeboden aan een methode die vervolgens de top 5 maakt. Voor het printen van de resultaten maken we gebuikt van pprint omdat dit het overzichtelijker print

We moesten ook een aantal rankings maken. We hebben er voor gekozen om per genre alle films die in dat genre thuis horen met daarbij alle rating op te slaan. Dit hebben we gedaan om het te maken om de gemiddelde rating per film te bereken en dan per genre een lijst te kunnen maken met daarin de top N films. Hiervoor gebruiken we de top 5 bepaal functionaliteit die we eerder gebruikt hebben

4.1 Resultaten

Hieronder de resultaten die wij uit onze algoritmes kregen. Wat opviel is dat drama in vele categorieën het hoogst scoort.

4.1.1 Top 5 filmgenre op basis van het geslacht

Categorie	1de	2de	3de	4de	5de
Vrouw(F)	Drama: 11008	Comedy: 8068	Romance 5858	Action 5442	Thriller 5086
Man(M)	Drama 28887	Comedy 21764	Action 20147	Thriller 16786	Romance 13603

Er zijn hier 2 zaken die opvallen namelijk de eerste twee genres zijn voor beide geslachten gelijk. Daarnaast valt op dat de beide geslachten dezelfde genres het vaakst beoordelen. Alleen staan deze in een andere volgorde

4.1.2 Top 5 filmgenre op basis van beroep

Categorie	1de	2de	3de	4de	5de
administrator	Drama 3099	Comedy 2203	Action 1858	Thriller 1570	Romance 1539
Artist	Drama 957	Comedy 617	Action 528	Thriller 476	Romance 462
Doctor	Drama 238	Comedy 168	Romance 133	Thriller 117	Action 110
Educator	Drama 4281	Comedy 2708	Romance 2006	Action 1962	Thriller 1767
Engineer	Drama 3153	Comedy 2438	Action 2277	Thriller 1712	Romance 1467
Entertainment	Drama 804	Comedy 574	Thriller 554	Action 499	Romance 342
Executive	Drama 1407	Comedy 951	Thriller 811	Action 808	Romance 592
Healthcare	Drama 1297	Comedy 722	Action 577	Thriller 563	Romance 499
Homemarker	Drama 104	Thriller 95	Comedy 93	Action 92	Romance 59
Lawyer	Drama 543	Comedy 460	Action 283	Romance 272	Thriller 247
Librarian	Drama 2557	Comedy 1576	Romance 1245	Action 988	Thriller 977
Marketing	Drama 842	Comedy 535	Action 477	Thriller 442	Romance 383
None	Action 301	Drama 268	Thriller 262	Comedy 255	Adventure 160
Other	Drama 4307	Comedy 3165	Action 3165	Action 2672	Romance 2117
Programmer	Drama 2800	Comedy 2418	Action 2322	Thriller 1750	Romance 1400

Retired	Drama 727	Comedy 512	Romance 321	Thriller 309	Action 278
Salesman	Drama 317	Comedy 314	Action 222	Thriller 200	Romance 168
Scientist	Drama 932	Action 566	Comedy 535	Thriller 469	Romance 440
Student	Drama 7777	Comedy 6958	Action 6398	Thriller 5130	Romance 4156
Technician	Drama 1295	Action 1079	Comedy 1056	Thriller 796	Romance 645
Writer	Drama 2208	Comedy 1574	Thriller 1304	Action 1292	Romance 1007

4.1.3 Top 5 op basis van Leeftijd categorie

Categorie	1de	2de	3de	4de	5de
0 -10	Comedy 19	Action 18	Adventure 14	Drama 12	Romance 12
10 - 20	Drama 2785	Comedy 2604	Action 2334	Thriller 1987	Romance 1453
20 - 30	Drama 14818	Comedy 12120	Action 10974	Thriller 9058	Romance 7500
30 -40	Drama 10275	Comedy 7729	Action 6424	Thriller 5047	Romance 5091
40 - 50	Drama 6463	Comedy 4296	Action 3536	Thriller 3096	Romance 3074
50 - 60	Drama 4166	Comedy 2247	Romance 1816	Thriller 1780	Action 1757
60 - 70	Drama 1288	Comedy 734	Romance 521	Thriller 493	Action 452
70 - 80	Drama 88	Comedy 83	Romance 53	Thriller 42	Action 35

4.1.4 Top 5 op basis van regio

We hebben gekeken naar hoe het zip code systeem in elkaar steekt. Door te kijken naar het eerste karakter kunnen we aardig regio's vormen

Regio	1de	2de	3de	4de	5de
0	Drama: 3815	Comedy: 2787	Action 2377	Thriller 1935	Romance 1817
1	Drama 4303	Comedy 3108	Action 2345	Thriller 2054	Romance 2005
2	Drama 4296	Comedy 3111	Action 2696	Thriller 2288	Romance 2032
3	Drama 2390	Comedy 1681	Action 1510	Thriller 1436	Romance 1159
4	Drama 2982	Comedy 2504	Action 2156	Thriller 1735	Romance 1596
5	Drama 4417	Comedy 3594	Action 3215	Thriller 2617	Romance 2326

6	Drama 3600	Comedy 2791	Action 2182	Thriller 2027	Romance 1812
7	Drama 2756	Comedy 1912	Action 1821	Thriller 1510	Romance 1396
8	Drama 2656	Comedy 2498	Action 2099	Thriller 1628	Romance 1331
9	Drama 7815	Comedy 5281	Action 4659	Thriller 4141	Romance 3577
E	Drama 221	Comedy 118	Action 112	Thriller 89	Romance 71
K	Comedy 15	Action 14	Sci-Fi 12	Romance 10	Adventure 9
L	Comedy 81	Drama 54	Action 49	Romance 40	Adventure 38
M	Drama 27	Action 17	Romance 13	Thriller 12	Sci-Fi 10
N	Drama 174	Action 158	Comedy 121	Thriller 112	Romance 86
R	Romance 23	Comedy 21	Drama 15	Thriller 8	Action 5
T	Drama 85	Comedy 48	Thriller 45	Action 30	Romance 28
V	Drama 183	Comedy 146	Thriller 134	Action 115	Romance 97
Y	Drama 98	Comedy 39	Thriller 37	Action 29	Romance 28

4.1.5 Top 10 movies

Om de 10 beste films te kiezen hebben we ervoor gekozen om het gemiddelde cijfer van de rankings van een film te kiezen. Dit heeft echter als nadeel dat wanneer een film 1 beoordeling heeft van 5 dat deze dan automatisch heel hoog staat. Ook hebben we geen rekening gehouden dat niet iedereen op dezelfde manier het cijfer bepaald. Dit zouden aanbevelingen kunnen zijn om het systeem te verbeteren.

Positie	Film title	Ranking	IMDB cijfer
1	They Made Me a Criminal (1939)	5.0	-
2	Santa with Muscles (1996)	5.0	-
3	Someone Else's America (1995)	5.0	-
4	Saint of Fort Washington, The (1993)	5.0	-
5	Entertaining Angels: The Dorothy Day Story (1996)	5.0	6.4
6	Marlene Dietrich: Shadow and Light	5.0	8.9
7	Star Kid (1997)	5.0	5.3
8	Aiqing wansui (1994)	5.0	-
9	Prefontaine (1997)	5.0	6.8
10	Great Day in Harlem, A	5.0	7,5

4.1.6 Ranking per genre

We hebben hier de films per genre gegroepeerd en vervolgens van elke film in het genre het gemiddelde berekend en vervolgens daar een top 5 van gemaakt

Genre	1ste	2 de	3 de	4 de	5 de
Action	Star wars 4.35	Godfather, The 4.28	Raiders of the Lost Ark 4.25	Titanic(1977) 4.24	Empire Strikes Back, The 4.20
Adventure	Star kid 5.0	Star wars 4.35	Raiders of the Lost Ark 4.25	Lawrence of Arabia 4.23	Empire Strikes Back, The 4.20
Animation	Close Shave, A (1995) 4.49	'Wrong Trousers, The (1993) 4.46	'Wallace & Gromit: The Best of Aardman 4.44	Faust(1994) 4.2	Grand day out a (1992) 4
Childeren	Star kid 5.0	Wizzard of Ozz the 4.07	Babe (1995) 3.99	Fly Away Home 3,90	Toy Story 3,87
Comedy	Santa with Muscles 5.0	Close Shave, A 4.49	'Wrong Trousers, The 4.44	North by Northwest 4.28	Shall we dance? 4.26
Crime	'They Made Me a Criminal 5.0	Usual Suspects, The 4.38	'Letter From Death Row, A 4.33	Godfather, The 4.28	'Crossfire 4.25
Documentary	Marlene Dietrich: Shadow and Light 5.0	Great Day in Harlem, A 5.0	Everest 4.5	'Maya Lin: A Strong Clear Vision 4.5	Hoop Dreams 4.09
Drama	Someone Else's America 5.0	Aiqing wansui 5.0	They Made Me a Criminal 5.0	Saint of Fort Washington, The 5.0	Entertaining Angels: The Dorothy Day 5.0
Fantasy	Star kid 5.0	E.T. the Extra-Terrestrial 3.83	Heavenly Creatures 3.67	20,000 Leagues Under the 3.5	Jumanji 3.31
Film-Noir	Manchurian Candidate, 4.259	Crossfire 4.25	Maltese Falcon, The 4.210	Sunset Blvd. 4.20	L.A. Confidential 4.16
Horror	Psycho 4.1	Alien 4.034	Young Frankenstein 3.94	Braindead 3.85	Shining, The 3.82

Musical	Wizzard of Ozz the 4.07	Top hat 4.04	Damsel in Distress 4.0	Singin' in the Rain 3.99	This Is Spinal Tap 3.90
Mystery	Rear Window 4.38	Third Man, The 4.33	Vertigo 4.25	Maltese Falcon 4.21	Amadeus 4.16
Romance	Casablanca 4.45	Star wars 4.35	Titanic 4.24	Empire Strikes Back 4.20	Affair to Remember 4.19
Sci-Fi	Star kid 5.0	Star wars 4.35	Love the Bomb 4.25	Empire Strikes Back, The 4.20	Blade Runner 4.13
Thriller	Close Shave, A 4.49	Rear Window 4.387	Usual Suspects, The 4.385	third Man, The 4.33	Lawrence of Arabia 4.23
War	Schindler's List 4.46	Casablanca 4.45	Star Wars 4.35	'Dr. Strangelove or: How I Learned to .. 4.25	Lawrence of Arabia 4.23
Western	High Noon 4.10	Wild Bunch, The 4.02	Butch Cassidy and the sundance ... 3.949	Magnificent Seven, The 3.942	Once Upon a Time in the West 3.88
Unkown	unknown 3.4	Good morning 1.0	-	-	-

5 Deel 2 van de opdracht (aanbeveling)

Bij deze opdracht is het de bedoeling om kennis te maken met clustering, door middel van het DBscan en het Kmeans algoritme. Daarna aanbevelingen te maken voor bepaalde gebruikers door middel van collaborative filtering.

De volgende aanbevelingen moeten gedaan worden voor een specifieke gebruiker.

- Welke 5 gebruikers het meest op die gebruiker lijken.
- Welke 5 films je zou aanbevelen op basis van de waarderingen van de andere filmliefhebbers met een vergelijkbare smaak. Leidt daartoe op basis van de geregistreerde aanbevelingen, relevante aanbeveling voor de gebruiker af.
- Welke 5 films het best bij hem passen op basis van de filmwaarderingen, die hij zelf gegeven heeft.

Optioneel mag de laatste aanbeveling gedaan worden aan de hand van extra data uit een andere dataset of van IMDB.

5.1 Aanpak

We hebben deze opdracht als volgt aan gepakt. Om de dataset beter te kunnen begrijpen, zijn we begonnen om wat uit te proberen met DBscan en Kmeans. Het resultaat wat DBscan ons gaf hebben wij doorgegeven aan Kmeans. Dit hebben wij gedaan, omdat DBscan een indicatie geeft van hoeveel clusters er in een dataset zitten. Kmeans moet daarentegen weten hoeveel clusters er zijn. Deze twee zijn daarom gemakkelijk te koppelen.

Hierna zijn we begonnen aan het maken van de aanbevelingen. Hiervoor hebben wij gebruik gemaakt van collaborative filtering. We zijn begonnen met het kiezen van een gebruiker. Dit hebben wij gedaan door de user te pakken met de grootste gelijkenis naar een andere gebruiker. Hiervan hebben wij dan de 5 gebruikers gepakt die het meeste hierop leken (Aanbeveling 1).

De volgende aanbeveling (aanbeveling 2) moest laten zien welke 5 films de gebruiker misschien zou willen zien. Dit zou moeten worden bepaald aan de hand van de andere gebruikers die veel leken op deze gebruiker. Voor deze aanbeveling hebben wij vijf verschillende strategieën toegepast.

De laatste aanbeveling is ons niet gelukt, aangezien de movies te weinig data hebben om content based recommendation op toe te passen. Wij hebben het geprobeerd om dit met collaborative filtering te achterhalen, maar als je de movies tegen elkaar uitzet in plaats van de users, krijg je een lijst met users terug in plaats van movies.

5.2 DBscan

Wij zijn begonnen met het toevoegen van DBscan. Dit zodat we een indicatie konden krijgen hoeveel clusters er ongeveer in onze dataset zitten. Om de afstand te kunnen bepalen van verschillende punten in onze dataset, hebben wij gebruik gemaakt van euclidean distance. Maar aangezien in onze dataset ook features bevat, die wij niet willen laten meetellen voor de afstand, hebben wij ervoor gekozen om onze eigen methode te schrijven voor het berekenen van de afstand. Hierbij hebben wij de euclidean distance alleen bepaald van alle genres, en namen we dingen zoals het user_id niet mee. Aangezien wij gebruik maken van de echte euclidean distance en DBscan normaal gesproken gebruik maakt van een schatting daarvan, is onze methode een stuk langzamer. Een leuke uitbreiding zal dan ook zijn om een schatting te berekenen in plaats van het echte antwoord.

Wij kwamen er al snel achter dat er geen tot weinig clusters in onze dataset zaten. We hebben veel verschillende dingen geprobeerd om tot een mooi antwoord te komen. Hierbij zijn we op zoek naar rond de negentien clusters, aangezien er negentien genres zijn. De volgende combinaties gaven de volgende uitkomsten:

	Epsilon			
Min_samples	30	35	40	45
5	0	120	660	822
6	0	75	618	813
7	0	32	513	795
8	0	18	483	786

Zoals je in de tabel hierboven kan zien, stijgt het aantal clusters als de epsilon omhoog wordt gezet. Dit valt als volgt te verklaren. De epsilon stelt de range van het clusteren voor. Dit is een range van een punt waar minimaal een aantal punten in moeten liggen, om mee geteld te mogen worden in dat cluster. Op een tweedimensionaal vlak is dit te zien als een cirkel. De reden, dat het aantal clusters groter wordt bij het verhogen van deze range, is dat er sneller een cluster gevonden kan worden als deze range groter wordt gemaakt. Dit kan echter het omgekeerde effect krijgen als deze range nog groter wordt. Uiteindelijk is de range zo groot dat alle punten worden meegerekend in het cluster, waardoor er nog maar één cluster overblijft. Met andere woorden, als je genoeg verschillende epsilons in een grafiek zou plotten, krijg je een soort parabool, die van 0 clusters naar ongeveer 850 clusters gaat en daarna weer afloopt naar 1 en dan een constante horizontale lijn krijgt.

Ín de tabel is ook te zien dat het aantal min_samples een klein effect heeft op de gegevens. Hoe lager dit aantal, des te meer clusters er gevonden worden. En als dit aantal hoger is, worden er juist minder clusters gevonden. Min_samples geeft aan hoe veel punten er in de range moeten zitten, voordat deze in het cluster opgenomen mogen worden. Als dit aantal lager is, kunnen punten sneller in een cluster worden opgenomen, waardoor er sneller een cluster kan worden gemaakt. Hierdoor ontstaan er meer clusters.

Volgens de tabel is de beste methode een combinatie van een epsilon van 35 en een min_samples van acht. Deze komt het meest overeen met onze schatting van negentien. Deze waarde geeft het aantal clusters weer dat in de dataset zouden moeten zitten. Daarom gaan wij gebruik maken van dit aantal in het Kmeans algoritme.

5.3 Kmeans

Na het toevoegen van DBscan zijn wij begonnen aan het Kmeans algoritme. Uit dit algoritme kunnen wij bepalen bij welke klasse een gebruiker valt. Bij Kmeans maken wij gebruik van één parameter. De parameter n_clusters geeft aan hoeveel clusters Kmeans moet maken. We maken hiervan gebruik van de waarde uit DBscan.

We hebben onze dataset opgesplitst in twee datasets. Eén dataset als trainingsset en één dataset als testset. Deze testset bestaat uit twintig gebruikers. Hieronder staan vijf runs met exact dezelfde parameters en datasets.

- [10 10 10 10 10 10 10 10 10 10 15 15 15 15 10 10 15 15 10 10]
- [10 10 10 10 10 10 10 10 10 10 2 2 2 2 10 10 2 2 10 10]
- [17 17 17 17 17 17 17 17 17 17 7 7 7 7 17 17 7 7 17 17]
- [6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6]
- [7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7]

Zoals je kunt zien zijn de clusters die aangegeven worden telkens anders. Ook lijkt het erop dat er telkens maar één of twee clusters aangegeven worden. Dit kan komen doordat er misschien exact één of twee grote clusters zijn. Dit kan betekenen dat deze dataset niet geschikt is voor clustering. Maar dit kan ook betekenen dat er bijvoorbeeld twee hele grote clusters zijn die duidelijk en verschil tonen.

Om er echt achter te komen of er clusters zijn, moet er verder onderzoek gedaan worden naar de data. Wij geloven dat DBscan en Kmeans vooral gebruikt moet worden om aan te geven dat het mogelijk is dat er clusters zijn. Je kunt niet met 100% zekerheid zeggen welke clusters er zijn en waar de exacte plaats is. Deze algoritmes kunnen niet vertellen hoe correct ze zijn. Er moet altijd onderzoek gedaan worden naar de datasets om te zien hoe correct ze zijn. Deze algoritmes geven dus puur aan dat er mogelijkheid is tot een cluster en een mogelijke plek hiervoor.

5.4 Collaborative filtering

Door middel van collaborative filtering moesten er aanbevelingen gedaan worden voor een specifiek gebruiker. Hiervoor zijn wij begonnen met het implementeren van een methode die de pearson correlation bepaalt. Hiervan hebben wij toen de gebruiker gepakt met de hoogste correlatie. Dit waren gebruikers 407 en 897. Deze gebruikers hebben niets gemeen behalve dat ze hetzelfde geslacht hebben. 407 en de andere gebruikers zien er als volgt uit:

- 407: Tussen 20 en 30 jaar, engineer, mannelijk en postcode begint met 0.
- 897: Tussen 30 en 40 jaar, other, mannelijk en postcode begint met 3.
- 104: Tussen 20 en 30 jaar, student, mannelijk en postcode begint met 5.
- 625: Tussen 20 en 30 jaar, programmer, mannelijk en postcode begint met 2.
- 38: Tussen 20 en 30 jaar, other, vrouwelijk en postcode begint met 5.
- 855: Tussen 50 en 60 jaar, librarian, mannelijk en postcode begint met 0.

Hierboven is te zien dat de meeste mensen hier tussen de 20 en de 30 jaar zitten. Ook is merendeel mannelijk en niet vrouwelijk. Het kan dus zijn dat hier een correlatie tussen zit.

Verder hebben wij vier verschillende strategieën gemaakt die een aanbeveling geven aan deze gebruiker aan de hand van de andere gebruikers. Ten eerste hebben we naar alle ratings gekeken van alle films van deze vijf gebruikers. Hierbij hebben we voor iedere film bijgehouden hoe vaak deze een beoordeling heeft gekregen van deze gebruikers. Hier kwam de volgende lijst uit:

- Men in Black (1997)
- Full Monty, The (1997)
- Starship Troopers (1997)
- Midnight in the Garden of Good and Evil (1997)
- Benny & Joon (1993)

Wij hebben hierbij Men in Black en Starship Troopers voor een gedeelte na gekeken in de data. Bij Starship Troopers kwamen wij erachter dat de gebruiker 407 (waar de aanbeveling voor is) deze film al gezien had. Dit is echter niet gewild. Daarom hebben wij een tweede strategie toegepast die alleen gaat kijken naar de films die gebruiker 407 niet gezien heeft. Hier kwam de volgende lijst uit:

- Men in Black (1997)
- Secrets & Lies (1996)
- Jackie Brown (1997)
- Hoodlum (1997)
- Mrs. Brown (Her Majesty, Mrs. Brown) (1997)

Wij hebben dit gecontroleerd door tussen de films te kijken die 407 al gezien heeft en deze stonden er dan ook niet tussen.

Er klopte nog iets niet aan het eerste algoritme. Er werd niet gekeken naar hoe hoog de rating was. Als Men in Black bijvoorbeeld alleen maar ratings van 1 kreeg, kon deze alsnog bovenaan staan. Daarom hebben wij een strategie bedacht die alleen de ratings hoger dan of gelijk aan 3 meetelt. Deze strategie geeft het volgende lijstje:

- Midnight in the Garden of Good and Evil (1997)
- Men in Black (1997)
- In & Out (1997)
- G.I. Jane (1997)
- Starship Troopers (1997)

Zoals je ziet is Men in Black één plaats omlaag gegaan. Dit houdt in dat dit dus in dat het voor deze film niet zo veel uitmaakt of er naar zijn ratings wordt gekeken. The Full monthy is daarentegen volledig uit de lijst. Dit houdt in dat deze film wel vaak gerate is maar de meerderheid daarvan lage ratings, binnen deze gebruikersgroep. Midnight in the Garden of Good and Evil heeft daarentegen juist weer veel goede ratings.

De vierde strategie die wij hebben toegepast kijken we niet naar de rating naar wie de rating heeft gedaan. Zo geeft de eerste gebruiker meer punten dan de vijfde gebruiker. In deze strategie is er verder niet gekeken of de film al gezien was of wat voor rating er gegeven werd. Deze strategie gaf de volgende lijst:

- Men in Black (1997)
- Full Monty, The (1997)
- Starship Troopers (1997)
- Midnight in the Garden of Good and Evil (1997)
- Benny & Joon (1993)

Zoals je ziet geeft deze lijst totaal geen verandering ten opzichte van de eerste strategie. Toch hebben wij hetzelfde geprobeerd door wel naar de rating te kijken. Hierbij kregen wij wel een verschil in de top 5.

Als laatste hebben we de strategieën 2 t/m 4 gecombineerd. Dit houdt in dat we keken naar welke films de gebruiker al gezien had, welke ratings de andere gebruikers aan deze films gaven en welke gebruikers de ratings gaven. Dit gaf de volgende top 5:

- Men in Black (1997)
- Secrets & Lies (1996)
- Hoodlum (1997)
- Jackie Brown (1997)
- Jackal, The (1997)

Men in black is dus de beste aanrader voor onze gebruiker. Hij is nog niet gezien en is goed gerate door de andere gebruikers.