1. Data

Data yang digunakan adalah data *movie rating* dari partisipan kelas *data mining* program magister Teknik Industri UGM. Data tersebut berisi mengenai preferensi partisipan dalam film-film yang ditampilkan. Film-film tersebut adalah; Toy Story 3 (2010) Inception (2010), Skyfall (2012), Her (2013), Parasite (2019), A Star Is Born (2018), Marriage Story (2019), Gone Girl (2014), The Nice Guys (2016), Avengers: End Game (2019), dan Train to Busan (2016). Partisipan mengisi preferensi film-film tersebut menggunakan *likert-scale* dari angka 1 (sangat tidak menyukai) hingga 5 (sangat menyukai). Jika partisipan belum menonton salah satu dari film tersebut, partisipan diberikan opsi angka 0 (belum menonton). Gambar 1 adalah sampel 10 data teratas data *movie rating*.

	Toy Story 3 (2010)	Inception (2010)	Skyfall (2012)	Her (2013)		A Star Is Born (2018)	Marriage Story (2019)	Gone Girl (2014)	The Nice Guys (2016)	Avengers: End Game (2019)	Train to Busan (2016)
USERNAME											
adind	0.0	5.0	4.0	0.0	0.0	4.0	0.0	5.0	4.0	5.0	0.0
adji	4.0	5.0	4.0	3.0	5.0	4.0	0.0	0.0	0.0	5.0	0.0
agrah	3.0	0.0	0.0	0.0	5.0	0.0	0.0	5.0	2.0	5.0	5.0
ajeng	4.0	0.0	0.0	0.0	0.0	5.0	0.0	5.0	0.0	4.0	5.0
alfin	0.0	0.0	5.0	0.0	4.0	4.0	0.0	0.0	0.0	5.0	0.0

Gambar 1. Sampel Data Movie Rating

2. Alat, Metode, dan Tahapan Penelitian

2.1 Alat Penelitian

Penelitian ini menggunakan *software* Python. Dalam pengerjaan penelitian ini, beberapa *library* Python digunakan untuk melakukan analisis data. Berikut merupakan daftar *library* Python yang digunakan pada penelitian ini.

1. Pandas

Pada penelitian ini, Pandas digunakan untuk melakukan modifikasi data hasil analisis.

2. Numpy

Pada penelitian ini, Numpy digunakan untuk melakukan komputasi saintifik pada data.

3. Matplotlib

Pada penelitian ini, Matplotlib digunakan untuk melakukan visualisasi hasil analisis segmentasi pasar.

4. Pyspark (pyspark.ml.recommendation ALS)

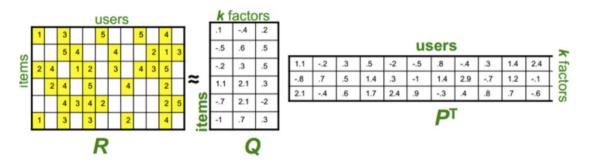
Pada penelitian ini Pyspark digunakan karena ingin menggunakan secara spesifik metode *Alternating Least Squares* (ALS) yang ada pada *library* Pyspark. Penjelasan mengenai metode ALS akan dipaparkan pada bagian metode penelitian.

5. Scikit-learn

Pada penelitian, Scikit-learn (Sklearn) digunakan untuk melakukan klusterisasi untuk segmentasi pasar menggunakan metode Kmeans.

2.2 Metode Penelitian

Penelitian ini menggunakan metode *Alternating Least Squares* (ALS) untuk melakukan segmentasi pasar dan membuat sistem rekomendasi. ALS adalah pengaturan faktorisasi matriks dengan rank rendah (Bell *et al.*, 2007). Menurut Zhou *et al.* (2018), ALS dapat melakukan algoritma paralel dengan akurasi yang baik. Dalam rekomendasi sistem, input ALS adalah matriks *movie rating yang* didekomposisi menjadi dua matriks produk yaitu matriks *user* dan *movie*. Dengan dilakukannya faktorisasi matriks, sel-sel kosong yang ada pada matriks awal dapat terisi dengan kalkulasi yang ada. Untuk memperjeleas konsep ini, Gambar 2 merupakan ilustrasi faktorisasi matriks.



Gambar 2. Ilustrasi Faktorisasi Matriks

Pada Gambar 2 dapat terlihat bahwa matriks produk (Q dan P^T) dapat mengisi sel-sel kosong pada matriks awal (R). Hasil kalkulasi pengisian sel-sel kosong tersebut yang dijadikan acuan untuk memberikan rekomendasi kepada *user*. Semakin tinggi sebuah nilai sel, semakin direkomendasikan film tersebut kepada *user*. ALS juga dapat mengungkapkan *latent features* yang ada pada data. *Latent features* adalah variabel yang membagi data menjadi beberapa bagian secara

natural. Pada penelitian ini, *latent features* tidak digunakan untuk melakukan segmentasi pasar.

Pada penerapan metode ALS pada *library* Pyspark, beberapa parameter dimasukkan seperti jumlah iterasi, parameter regulasi, dan jumlah rank. Selain itu, evaluasi terhadap model ALS dapat dilakukan dengan menghitung error menggunakan *root mean square error* (RMSE). Pada penelitian ini, penentuan parameter akan dilakukan secara iteratif hinggan *stopping criteria* RMSE didapatkan. Jika *stopping criteria* terpenuhi, model kemudian digunakan untuk merekomendasikan film kepada masing-masing user.

Hasil rekomendasi tersebut juga digunakan untuk melakukan klusterisasi *user*. Klusterisasi dilakukan menggunakan metode K-means. Sebelum dilakukan klusterisasi menggunakan metode K-means, terlebih dahulu dilakukan reduksi variabel menggunakan metode *Principal Component Analysis* (PCA).

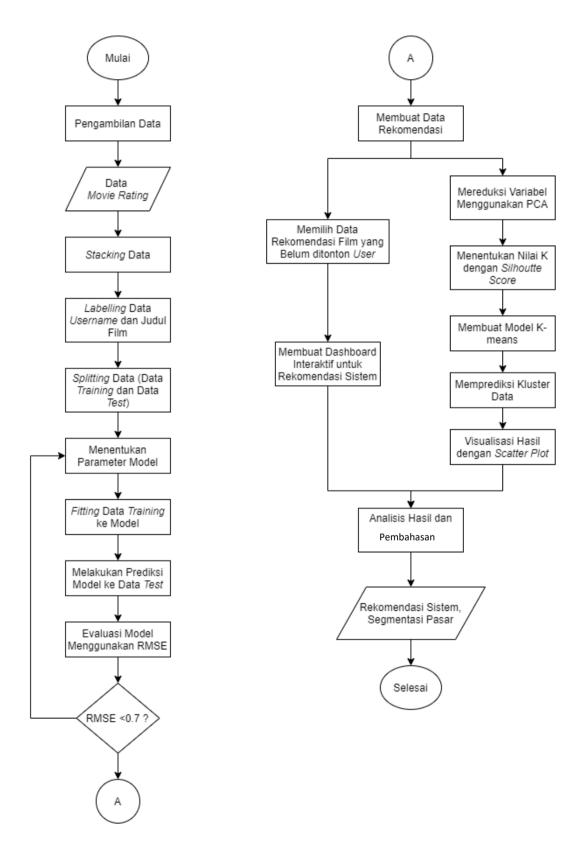
PCA adalah teknik mereduksi dimensi atau variabel sebuah dataset untuk meningkatkan interpretasi data tersebut tanpa mengurangi infromasi dari data tersebut (Jolliffe dan Cadima, 2016). Pada penelitian ini, PCA dilakukan untuk mereduksi jumlah film menjadi dua variabel saja. Sementara itu, K-means adalah salah satu metode klusterisasi dengan melakukan pengukuran jarak antara masingmasing individu data terhadap titik pusat atau biasa disebut *centroid*. Menurut (Tan *et al.*, 2005), berikut merupakan tahapan K-means *clustering*.

- 1. Menentukan jumlah kelompok (K)
- 2. Menentukan *centroid* masing-masing kelompok
- 3. Mengelompokkan data sehingga terbagi menjadi K kelompok berdasarkan kedekatan data yang diobservasi terhadap *centroid*
- 4. Melakukan pembaruan terhadap titik centroid
- 5. Ulangi tahap 2 hingga 4, sampai titik *centroid* tidak berubah

Pada penelitian ini, nilai K yang digunakan ditentukan berdasarkan nilai silhoutte score yang dihasilkan dari klusterisasi yang terbentuk. Silhoutte score adalah sebuah nilai yang mengukur apakah sebuah data individu berada di kluster yang tepat dibandingkan dengan kluster lainnya. Untuk memperjelas tahapan pengerjaan, Gambar 3 merupakan diagram alir penelitian ini.

2.3 Tahapan Penelitian

Tahapan penelitian ini dijelaskan pada gambar diagram alir berikut.



Gambar 3. Diagram Alir Penelitian

3. Hasil dan Pembahasan

3.1 Hasil dan Pembahasan Sistem Rekomendasi

Parameter akhir yang digunakan model ALS yaitu; jumlah iterasi adalah 15, parameter regulasi adalah 0,2 dan jumlah rank adalah 2. Sementara itu, RMSE akhir yang didapatkan adalah 0,687. Lampiran 1 adalah rekomendasi untuk film-film yang belum ditonton oleh user. Script pemrograman dari pemodelan data dan hasil RMSE yang didapatkan ditunjukkan pada Gambar 4. Penelitian ini juga menghasilkan sebuah output berbentuk dashboard interaktif yang dimana user melakukan input nama dan sistem rekomendasi memberikan output berupa rekomendasi film-film yang belum ditonton user dan perkiraan rating user dari sistem. Gambar 5 dan 6 merupakan ilustrasi dashboard interaktif yang dibuat.

Gambar 4. Script Pemrograman dan Nilai RMSE yang Didapatkan

```
Hello, welcome to IE Movie Recommendation System!
Please enter your name to get your movie recommendation
```

Gambar 5. Input Nama Dashboard Interaktif

```
Here are your movie recommendation and the rating prediction

USERNAME Movie Rating_Prediction

13 adji Train_to_Busan 4.46
15 adji Gone_Girl 4.35
19 adji Marriage_Story 3.81
20 adji The_Nice_Guys 3.44

Hope you like it and happy watching!
```

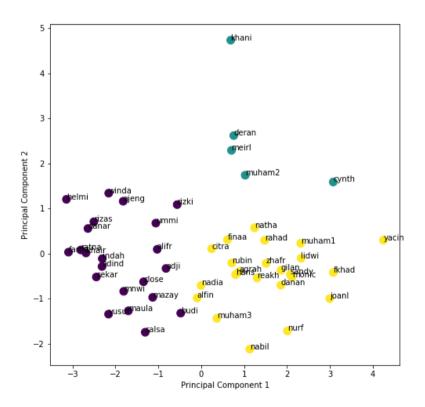
Gambar 6. Output Rekomendasi Film Dashboard Interaktif

Nilai RMSE 0,687 menjelaskan bahwa *output* dari model yang terbuat memiliki ratarata deviasi sebesar 0,687. Angka 0,687 bukan angka yang kecil dalam skala penilaian 1 – 5. Namun demikian, angka tersebut sudah menjadi salah satu angka error terendah yang berhasil didapatkan. Penyebab hal tersebut adalah sedikitnya jumlah data yang tersedia. Dengan hanya 56 *user* dan 11 film, informasi-informasi yang dibutuhkan mungkin masih

belum tercukupi. Di sisi lain, konsep *faktorisasi matriks* yang digunakan pada metode ALS memiliki akurasi yang tinggi. Perbandingan hasil rekomendasi yang didapatkan perlu dibandingkan dengan metode-metode lainnya untuk membuktikan bahwa hasil yang didapatkan sudah optimal. Selain itu, validasi hasil kepada *user* juga diperlukan untuk pengembangan model rekomendasi yang digunakan.

3.2 Hasil dan Pembahasan Segmentasi Pasar

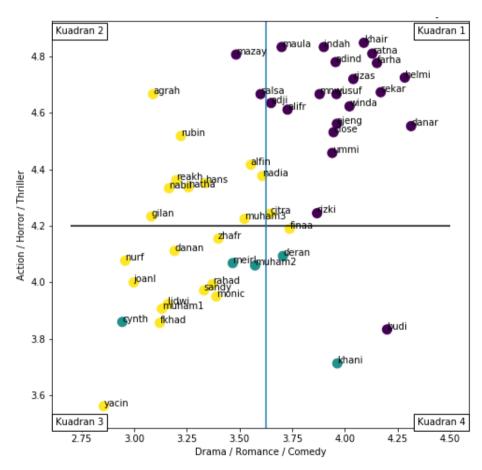
Segmentasi pasar pada penelitian ini dilakukan menggunakan data hasil rekomendasi. Data-data rekomendasi tersebut digunakan untuk mengisi sel-sel kosong yang ada pada data *movie rating* awal. Pada penelitian ini, sebelum masuk klusterisasi, langkah pertama yang dilakukan adalah mereduksi variabel menggunakan PCA. Jumlah variabel yang ditentukan adalah 2 agar bisa dilakukan visualisasi hasil. Kemudian dilakukan perhitungan *silhouette score* untuk menentukan jumlah kluster yang terbentuk. Jumlah kluster dengan nilai terbaik adalah 3 kluster. Kemudian dilakukan pemodelan data menggunakan metode k-means. Gambar 7 merupakan visualisasi hasil klusterisasi.



Gambar 7. Visualisasi Hasil Klusterisasi

Gambar 7 menunjukkan visualisasi hasil klusterisasi. Dapat terlihat bahwa terdapat 3 kluster dengan 2 kluster besar. Namun demikian, visualisasi tersebut belum dapat menggambarkan segmentasi pasar yang ada.

Untuk mendapatkan segmentasi pasar yang ada, penelitian ini membagi film-film yang ada kedalam dua kelompok. Kelompok pertama adalah kelompok film-film santai dan menghibur. Anggota kelompok ini adalah film-film bergenre *drama/romance/comedy* yaitu Toy Story 3 (2010), Her (2013), A Star Is Born (2018), Marriage Story (2019), dan The Nice Guys (2016). Kelompok kedua adalah kelompok film-film menegangkan yang memacu adrenaline *user* atau penonton. Anggota kelompok ini adalah film-film bergenre *action/horror/thriller* yaitu film Inception (2010), Skyfall (2012), Parasite (2019), Gone Girl (2014), Avengers: End Game (2019), dan Train to Busan (2016). Film-film tersebut kemudian dipetakan berdasarkan nilai rata-rata rating *user* terhadap masing-masing kelompok. Gambar 8 merupakan visualisasi hasil modifikasi berdasarkan nilai rata-rata masing-masing kelompok dari *user*.



Gambar 8. Visualisasi Hasil Klusterisasi Modifikasi

Dari gambar tersebut dapat terlihat *user-user* terbagi kedalam beberapa kuadran. Anggota kluster 1 (warna kuning) mayoritas menempati kuadran 2 dan 3. Kuadran 2 dan 3 berkarakteristik memiliki nilai rata-rata kelompok film *drama/comedy/romance* yang

rendah. Namun demikian, karena tersebar pada kluster 2 dan 3, *user* yang berada pada kluster ini tidak memiliki kecendurungan terhadap kelompok film *action/horror/thriller*. Dari karakteristik tersebut dapat disimpulkan bahwa kemungkinan anggota kluster 1 memiliki kecenderungan tidak menyukai film santai dan menghibur bergenre *drama/comedy/romance*, namun tidak memiliki kecenderungan preferensi terhadap film menegangkan yang memacu adrenaline bergenre *action/horror/thriller*.

Untuk kluster 2 (warna ungu), posisi mayoritas anggotanya berada pada kuadran 1. Kuadran 1 berkarakteristik memiliki nilai rata-rata kelompok film *drama/comedy/romance* dan kelompok film *action/horror/thriller* yang tinggi. Dari karakteristik tersebut, dapat disimpulkan bahwa kemungkinan anggota kluster 2 menyukai kedua jenis kelompok film. Sementara itu, anggota kluster 3 (warna hijau) tersebar pada kuadran 3 dan 4. Kedua kuadran ini berkarakteristik memiliki rata-rata nilai kelompok film *action/horror/thriller* yang rendah. Namun demikian, karena tersebar dalam dua kluster tersebut, *user* yang berada pada kluster ini tidak memiliki kecendurungan terhadap kelompok film *drama/comedy/romance*. Dari karakteristik tersebut dapat disimpulkan bahwa kemungkinan anggota kluster 3 memiliki kecenderungan tidak menyukai film menegangkan yang memacu adrenaline bergenre *action/horror/thriller*, namun tidak memiliki kecenderungan preferensi terhadap film santai dan menghibur bergenre *drama/comedy/romance*.

3.3 Kelebihan dan Kekurangan Metode ALS dan K-means

Metode ALS menggunakan konsep *faktorisasi matriks* dalam pengerjaannya. Konsep ini terbilang cukup tepat untuk melakukan rekomendasi film karena konsep ini dapat mengukur sebuah nilai rating tanpa menggunakan data diluar data *user* dan *movie*. Selain itu, metode ini juga memiliki akurasi yang cukup baik, walau dalam studi kasus penelitian ini masih dibutuhkan perbandingan dengan metode lainnya. Namun demikian, metode ini juga memiliki beberapa kelemahan. Salah satu kelemahannya adalah membutuhkan iterasi yang cukup banyak untuk dapat menghasilkan hasil yang optimal. Walaupun memiliki akurasi yang baik, hasil optimal sangat sulit didapatkan dengan metode ini.

Metode k-means merupakan salah satu metode sederhana dalam klusterisasi. Namun demikian, metode ini memiliki performa yang baik. Selain itu, interpretasi dari penerapan metode ini juga dipahami. Metode k-means juga memiliki beberapa kelemahan. Salah satunya adalah metode ini sulit dalam menenetukan jumlah kluster. Dalam studi kasus pada penelitian ini, jumlah kluster yang terbentuk berdasarkan *silhoutte score* yang rendah.

Daftar Pustaka

- Bell, R., Koren, Y. and Volinsky, C., 2007. Modeling relationships at multiple scales to improve accuracy of large recommender systems. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining KDD '07*.
- Jolliffe, I. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).
- Tan, P., Steinbach, M., Karpatne, A. and Kumar, V., 2005. *Introduction To Data Mining*. 1st ed. Boston: Addison-Wesley Longman Publishing Co., Inc.
- Zhou, Y., Wilkinson, D., Schreiber, R. and Pan, R., 2008. Large-Scale Parallel Collaborative Filtering for the Netflix Prize. *Algorithmic Aspects in Information and Management*, pp.337-348.

LAMPIRAN

Lampiran 1. Rekomendasi Film-film yang Belum Ditonton Oleh User

USERNAME	Movie	Rating_Prediction
adind	Parasite	4.9
adind	Train_to_Busan	4.74
adind	Toy_Story	4.25
adind	Marriage_Story	4.05
adind	Her	3.45
adji	Train_to_Busan	4.44
adji	Gone_Girl	4.32
adji	Marriage_Story	3.8
adji	The_Nice_Guys	3.42
agrah	Inception	4.14
agrah	Skyfall	3.83
agrah	A_Star_Is_Born	3.68
agrah	Marriage_Story	3.63
agrah	Her	3.09
ajeng	Parasite	4.74
ajeng	Inception	4.46
ajeng	Skyfall	4.12
ajeng	Marriage_Story	3.91
ajeng	The_Nice_Guys	3.52
ajeng	Her	3.33
alfin	Train_to_Busan	4.23
alfin	Inception	4.12
alfin	Gone_Girl	4.11
alfin	Toy_Story	3.79
alfin	Marriage_Story	3.61
alfin	The_Nice_Guys	3.25
alfin	Her	3.07
alifr	Parasite	4.69
alifr	Inception	4.41
alifr	Gone_Girl	4.41
alifr	Skyfall	4.08
alifr	A_Star_Is_Born	3.93
alifr	Marriage_Story	3.87
alifr	The_Nice_Guys	3.49
alifr	Her	3.3
ashle	A_Star_Is_Born	2.86
ashle	Marriage_Story	2.81
citra	Train_to_Busan	4.26
citra	Inception	4.15

citra	Gone_Girl	4.15
citra	Skyfall	3.84
citra	Toy_Story	3.82
citra	The_Nice_Guys	3.28
citra	Her	3.1
close	Parasite	4.72
close	Gone_Girl	4.44
close	A_Star_Is_Born	3.95
close	Marriage_Story	3.9
close	The_Nice_Guys	3.51
close	Her	3.32
cynth	Gone_Girl	3.7
cynth	Skyfall	3.43
cynth	The_Nice_Guys	2.93
cynth	Her	2.77
danan	Parasite	4.16
danan	Train_to_Busan	4.01
danan	Inception	3.91
danan	Gone_Girl	3.91
danan	Skyfall	3.62
danan	A_Star_Is_Born	3.48
danan	Marriage_Story	3.43
danan	The_Nice_Guys	3.09
danan	Her	2.92
danar	Train_to_Busan	4.7
danar	Gone_Girl	4.58
danar	A_Star_Is_Born	4.08
danar	Marriage_Story	4.02
danar	Her	3.42
deran	Inception	3.91
deran	Skyfall	3.62
deran	A_Star_Is_Born	3.48
deran	The_Nice_Guys	3.09
deran	Her	2.92
farha	Inception	4.64
farha	Marriage_Story	4.07
farha	The_Nice_Guys	3.66
finaa	Train_to_Busan	4.1
finaa	Inception	4
finaa	Gone_Girl	3.99
finaa	Marriage_Story	3.51
finaa	The_Nice_Guys	3.16
finaa	Her	2.99
fkhad	Inception	3.55
fkhad	Gone_Girl	3.55

fkhad	Marriage_Story	3.12
fkhad	The_Nice_Guys	2.81
fkhad	Her	2.65
gilan	Inception	3.88
gilan	Gone_Girl	3.88
gilan	Skyfall	3.59
gilan	Marriage_Story	3.41
gilan	The_Nice_Guys	3.07
gilan	Her	2.9
grita	Train_to_Busan	2.44
grita	Skyfall	2.2
hans	Inception	4.13
hans	Gone_Girl	4.12
hans	Skyfall	3.82
hans	A_Star_Is_Born	3.68
hans	Marriage_Story	3.62
hans	The_Nice_Guys	3.26
hans	Her	3.08
hayya	Parasite	2.93
hayya	Gone_Girl	2.75
hayya	A_Star_Is_Born	2.45
hayya	Marriage_Story	2.42
hayya	The_Nice_Guys	2.18
helmi	Parasite	5.01
helmi	Train_to_Busan	4.84
helmi	Inception	4.71
helmi	Gone_Girl	4.71
helmi	Marriage_Story	4.13
helmi	The_Nice_Guys	3.72
helmi	Her	3.52
indah	Marriage_Story	4.03
indah	Her	3.43
joanl	The_Nice_Guys	2.98
khair	Gone_Girl	4.7
khair	Skyfall	4.35
khair	A_Star_Is_Born	4.19
khair	The_Nice_Guys	3.72
khair	Her	3.51
khani	Inception	3.77
khani	Skyfall	3.48
khani	Her	2.81
lidwi	Inception	3.76
lidwi	Gone_Girl	3.76
lidwi	The_Nice_Guys	2.97
lidwi	Her	2.81

maula	A_Star_Is_Born	3.96
maula	The_Nice_Guys	3.51
mazay	Train_to_Busan	4.46
mazay	Inception	4.34
mazay	A_Star_Is_Born	3.87
mazay	Marriage_Story	3.81
mazay	The_Nice_Guys	3.43
mazay	Her	3.25
meirl	Gone Girl	3.83
meirl	Skyfall	3.55
meirl	A_Star_Is_Born	3.41
meirl	The_Nice_Guys	3.03
meirl	Her	2.86
mnwi	Marriage_Story	3.99
mnwi	Her	3.39
monic	Gone_Girl	3.69
monic	Marriage_Story	3.24
monic	The_Nice_Guys	2.92
monic	Her	2.76
muham1	Parasite	4.06
muham1	Avengers_End_Game	4.02
muham1	Train_to_Busan	3.92
muham1	Inception	3.82
muham1	Skyfall	3.53
muham1	A_Star_Is_Born	3.4
muham1	Marriage_Story	3.35
muham1	The_Nice_Guys	3.02
muham1	Her	2.85
muham2	Avengers_End_Game	4.21
muham2	Train_to_Busan	4.1
muham2	Gone_Girl	3.99
muham2	Toy_Story	3.68
muham2	The_Nice_Guys	3.16
muham2	Her	2.99
muham3	Train_to_Busan	4.21
muham3	Gone_Girl	4.1
muham3	A_Star_Is_Born	3.65
muham3	Marriage_Story	3.6
muham3	The_Nice_Guys	3.24
muham3	Her	3.06
nabil	Gone_Girl	3.98
nabil	Toy_Story	3.67
nabil	The_Nice_Guys	3.15
nabil	Her	2.98
nadia	Train_to_Busan	4.17

nadia	Gone_Girl	4.06
nadia	Her	3.03
natha	Gone_Girl	4
natha	A_Star_Is_Born	3.57
natha	Marriage_Story	3.51
natha	The_Nice_Guys	3.17
natha	Her	2.99
nurf	Avengers_End_Game	3.95
nurf	Skyfall	3.47
nurf	The_Nice_Guys	2.96
nurf	Her	2.8
rahad	Parasite	4.08
rahad	Inception	3.84
rahad	A_Star_Is_Born	3.42
rahad	Marriage_Story	3.37
rahad	The_Nice_Guys	3.04
ratna	Avengers_End_Game	4.84
ratna	The_Nice_Guys	3.63
reakh	Train_to_Busan	4.15
rizas	Train_to_Busan	4.71
rizas	Inception	4.58
rizas	A_Star_Is_Born	4.08
rizas	Marriage_Story	4.02
rizas	The_Nice_Guys	3.62
rizas	Her	3.42
rizki	Parasite	4.54
rizki	Train_to_Busan	4.38
rizki	Inception	4.27
rizki	Gone_Girl	4.26
rizki	Skyfall	3.95
rizki	Marriage_Story	3.74
rizki	The_Nice_Guys	3.37
rizki	Her	3.19
rubin	Inception	4.1
rubin	Toy_Story	3.77
rubin	The_Nice_Guys	3.24
rubin	Her	3.06
sandy	Gone_Girl	3.82
sandy	A_Star_Is_Born	3.4
sandy	Marriage_Story	3.35
sandy	The_Nice_Guys	3.02
sandy	Her	2.85
sekar	Inception	4.78
sekar	Gone_Girl	4.78
sekar	Skyfall	4.42

sekar	A_Star_Is_Born	4.26
sekar	Marriage_Story	4.2
sekar	The_Nice_Guys	3.78
sekar	Her	3.57
ummi	Avengers_End_Game	4.65
ummi	Train_to_Busan	4.53
ummi	Inception	4.41
ummi	Skyfall	4.08
ummi	Marriage_Story	3.87
ummi	The_Nice_Guys	3.49
ummi	Her	3.3
winda	Train_to_Busan	4.72
winda	A_Star_Is_Born	4.09
yacin	Train_to_Busan	3.5
yacin	Inception	3.41
yacin	Gone_Girl	3.41
yacin	Marriage_Story	2.99
yacin	The_Nice_Guys	2.7
yacin	Her	2.55
yusuf	A_Star_Is_Born	3.97
yusuf	Marriage_Story	3.91
yusuf	The_Nice_Guys	3.53
yusuf	Her	3.33
zhafr	Gone_Girl	3.91