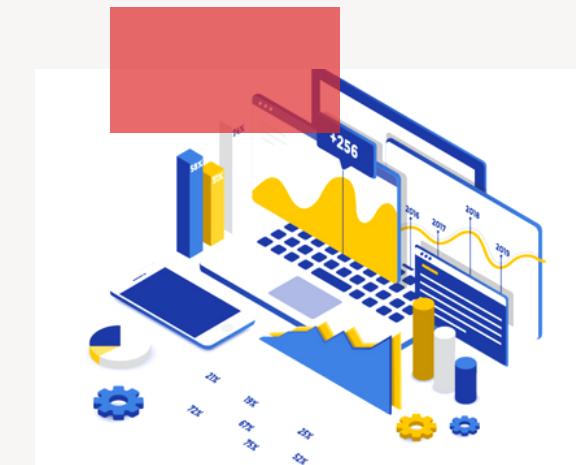


Movie Recommendation Engine

Capstone Project
By April Kwong



Contents



Problem

- Business Problem
- Stakeholder

Data

- Data Cleaning
- Feature Engineering

Model

- CountVectorizer & Cosine Similarity
- Movie Recommendations
- Next Steps





Movie Streaming Platforms

■ Stakeholder

Movie streaming platforms

■ Business Problem

To increase profit

■ Data Problem

Increase the time spent on the website by
recommending similar movies which users
have watched



Data

- Kaggle
 - Combined 2 datasets together
 - 4802 movies



Data

	id	title	cast	crew	genres	keywords
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "cre...	[{"credit_id": "52fe48009251416c750aca23", "de...	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Ad...	[{"id": 1463, "name": "culture clash"}, {"id": 1464, "name": "...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Sparrow", "cre...	[{"credit_id": "52fe4232c3a36847f800b579", "de...	[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "...	[{"id": 270, "name": "ocean"}, {"id": 726, "name": "...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cre...	[{"credit_id": "54805967c3a36829b5002c41", "de...	[{"id": 28, "name": "Action"}, {"id": 12, "name": "...	[{"id": 470, "name": "spy"}, {"id": 818, "name": "...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Batman", "cre...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...	[{"id": 28, "name": "Action"}, {"id": 80, "name": "...	[{"id": 849, "name": "dc comics"}, {"id": 853, "name": "...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "cre...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...	[{"id": 28, "name": "Action"}, {"id": 12, "name": "...	[{"id": 818, "name": "based on novel"}, {"id": 1463, "name": "...

ID of the movie



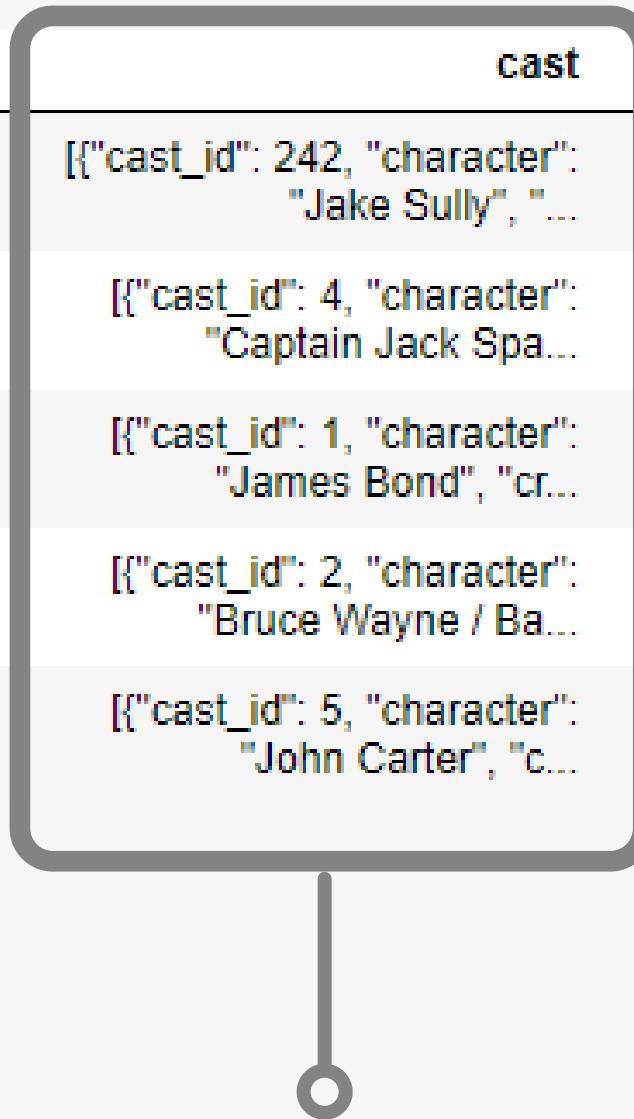
Data

			id	title	cast	crew	genres	keywords
0	19995		Avatar	["cast_id": 242, "character": "Jake Sully", "cre...	["credit_id": "52fe48009251416c750aca23", "de...	["id": 28, "name": "Action"], {"id": 12, "nam...	["id": 1463, "name": "culture clash"], {"id": ...	
1	285	Pirates of the Caribbean: At World's End		["cast_id": 4, "character": "Captain Jack Spa...	["credit_id": "52fe4232c3a36847f800b579", "de...	["id": 12, "name": "Adventure"], {"id": 14, "...	["id": 270, "name": "ocean"], {"id": 726, "na...	
2	206647		Spectre	["cast_id": 1, "character": "James Bond", "cr...	["credit_id": "54805967c3a36829b5002c41", "de...	["id": 28, "name": "Action"], {"id": 12, "nam...	["id": 470, "name": "spy"], {"id": 818, "name...	
3	49026	The Dark Knight Rises		["cast_id": 2, "character": "Bruce Wayne / Ba...	["credit_id": "52fe4781c3a36847f81398c3", "de...	["id": 28, "name": "Action"], {"id": 80, "nam...	["id": 849, "name": "dc comics"], {"id": 853, ...	
4	49529		John Carter	["cast_id": 5, "character": "John Carter", "c...	["credit_id": "52fe479ac3a36847f813eaa3", "de...	["id": 28, "name": "Action"], {"id": 12, "nam...	["id": 818, "name": "based on novel"], {"id": ...	

Title of the movie



Data



Full credits for cast with actor and actresses listed in the order they appear in the credits.

E.g. "cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a3685532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order": 0



Data

	id	title	cast	crew	genres	keywords
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "c...}	[{"credit_id": "52fe48009251416c750aca23", "de...}	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id": ...}
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...}	[{"credit_id": "52fe4232c3a36847f800b579", "de...}	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...}	[{"credit_id": "54805967c3a36829b5002c41", "de...}	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...}	[{"credit_id": "52fe4781c3a36847f81398c3", "de...}	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 849, "name": "dc comics"}, {"id": 853, ...}
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...}	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...}	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 818, "name": "based on novel"}, {"id": 12, ...}



Full credits for crew listed with details on "department" and "job".

E.g. "credit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"



Data

	id	title	cast	crew	genres	keywords
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "c...}	["52fe48009251416c750aca23", "de...]	[{"id": 28, "name": "Action"}, {"id": 12, "nam...]	[{"id": 1463, "name": "culture clash"}, {"id": ...]
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...}	["52fe4232c3a36847f800b579", "de...]	[{"id": 12, "name": "Adventure"}, {"id": 14, "...]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...]
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...}	["54805967c3a36829b5002c41", "de...]	[{"id": 28, "name": "Action"}, {"id": 12, "nam...]	[{"id": 470, "name": "spy"}, {"id": 818, "name...]
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...}	["52fe4781c3a36847f81398c3", "de...]	[{"id": 28, "name": "Action"}, {"id": 80, "nam...]	[{"id": 849, "name": "dc comics"}, {"id": 853,...]
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...}	["52fe479ac3a36847f813eaa3", "de...]	[{"id": 28, "name": "Action"}, {"id": 12, "nam...]	[{"id": 818, "name": "based on novel"}, {"id": ...]

Genres of the movie



Data

Keywords related to the movie



Data

	id	title	cast	crew	genres	keywords	director
0	19995	Avatar	[Sam Worthington, Zoe Saldana, Sigourney Weaver, ...]	['52fe48009251416c750aca23', 'de...']	[Action, Adventure, Fantasy]	[culture clash, future, space war]	James Cameron
1	285	Pirates of the Caribbean: At World's End	[Johnny Depp, Orlando Bloom, Keira Knightley, ...]	['52fe4232c3a36847f800b579', 'de...']	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island]	Gore Verbinski
2	206647	Spectre	[Daniel Craig, Christoph Waltz, Léa Seydoux, R...]	['54805967c3a36829b5002c41', 'de...']	[Action, Adventure, Crime]	[spy, based on novel, secret agent]	Sam Mendes
3	49026	The Dark Knight Rises	[Christian Bale, Michael Caine, Gary Oldman, A...]	['52fe4781c3a36847f81398c3', 'de...']	[Action, Crime, Drama]	[dc comics, crime fighter, terrorist]	Christopher Nolan
4	49529	John Carter	[Taylor Kitsch, Lynn Collins, Samantha Morton, ...]	['52fe479ac3a36847f813eaa3', 'de...']	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion]	Andrew Stanton

Extraction of Data

- Director's name from Crew column
- Main actors (top 5) from Cast
- Main genres (top 3) from Genres column
- Main keywords (top 3) from Keywords column



Data

	id	title	cast	genres	keywords	director
0	19995	Avatar	[samworthington, zoesaldana, sigourneyweaver, ...]	[action, adventure, fantasy]	[cultureclash, future, spacewar]	jamescameron
1	285	Pirates of the Caribbean: At World's End	[johnnydepp, orlandobloom, keiraknightley, ste...]	[adventure, fantasy, action]	[ocean, drugabuse, exoticisland]	goreverbinski
2	206647	Spectre	[danielcraig, christophwaltz, l��aseydoux, ralp...]	[action, adventure, crime]	[spy, basedonnovel, secretagent]	sammendes
3	49026	The Dark Knight Rises	[christianbale, michaelcaine, garyoldman, anne...]	[action, crime, drama]	[dccomics, crimefighter, terrorist]	christophernolan
4	49529	John Carter	[taylorkitsch, lynncollins, samanthamorton, wi...]	[action, adventure, sciencefiction]	[basedonnovel, mars, medallion]	andrewstanton

Extraction of Data

- Director's name from Crew column
- Main actors (top 5) from Cast
- Main genres (top 3) from Genres column
- Main keywords (top 3) from Keywords column

Data Cleaning

- Checked for null values
- Checked for duplicate titles
- Removed spaces
- Converted to lower case
- Dropped unrequired columns



Data

Preparation of Data

- Combined the data into 1 column

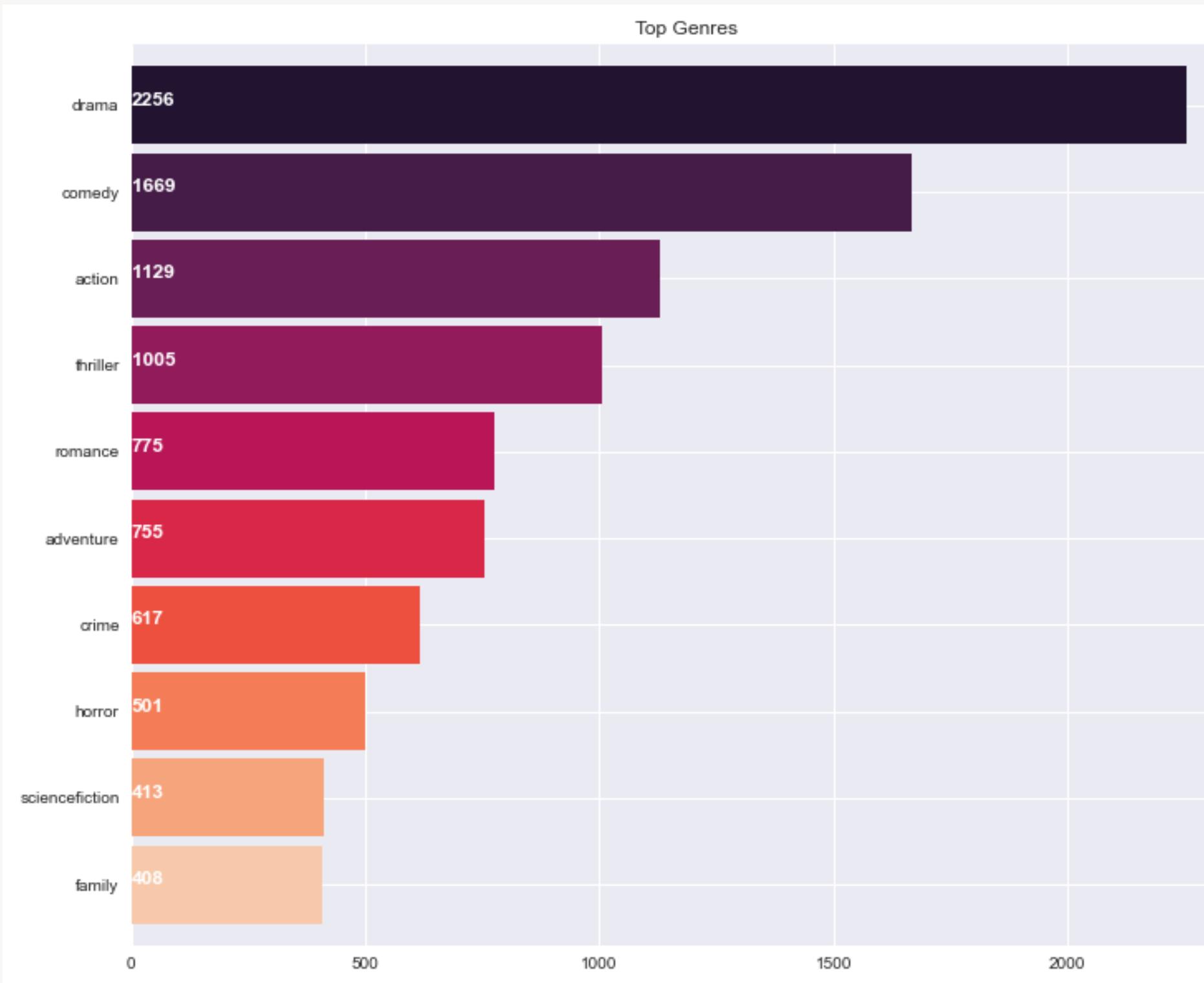
'action adventure fantasy cultureclash future
spacewar samwashington zoesaldana
sigourneyweaver stephenlang
michellerodriguez jamescameron'



Exploratory Data Analysis



Top 10 Genres



■ Drama: 2265

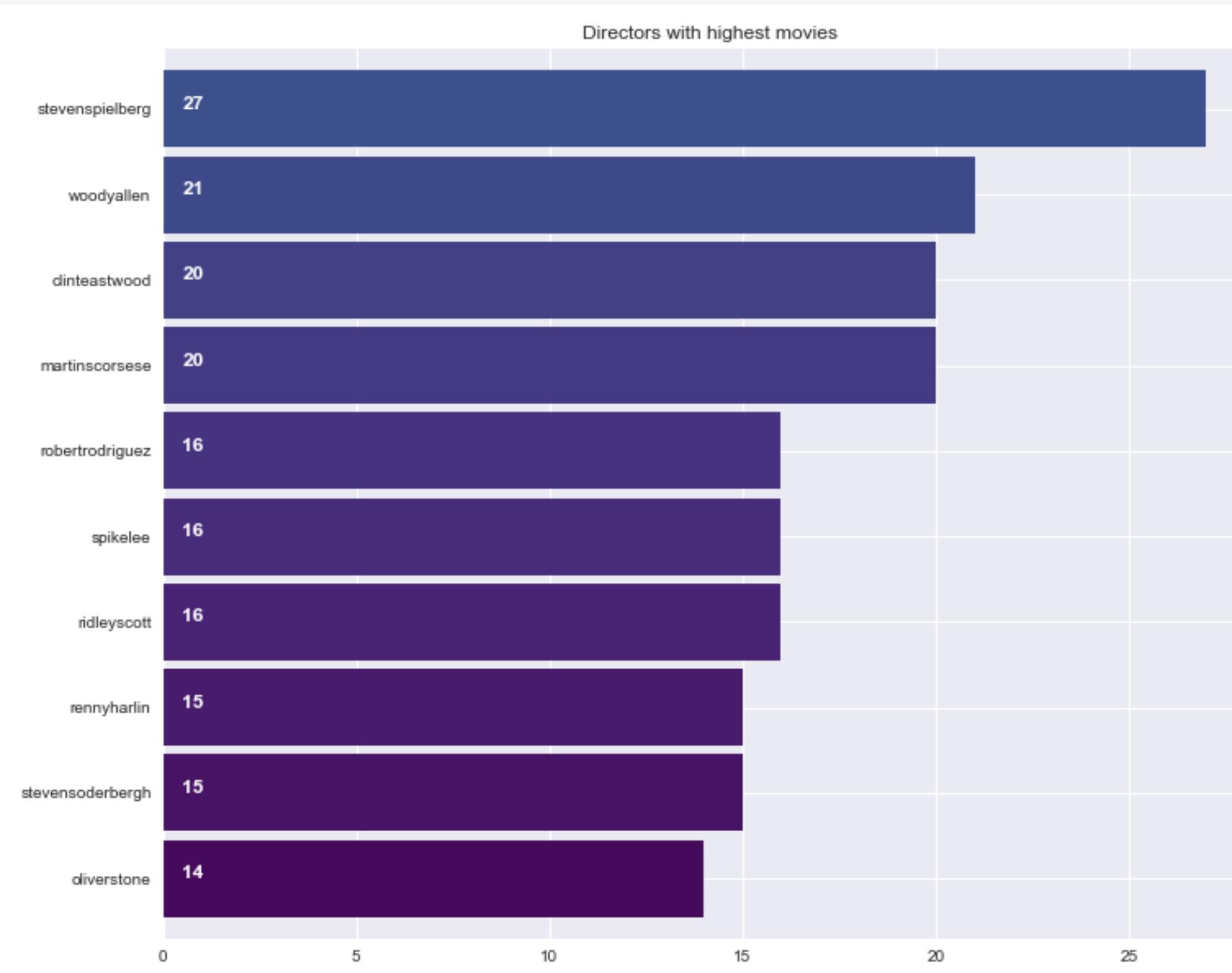
■ Comedy: 1669

■ Action: 1129



Top 10 Directors

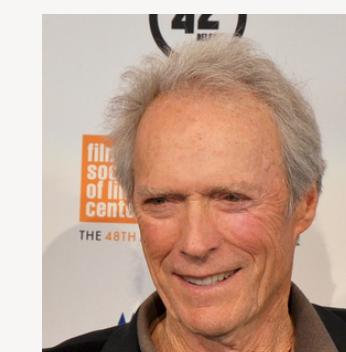
(who directed the most number of films)



Steven Spielberg: 27



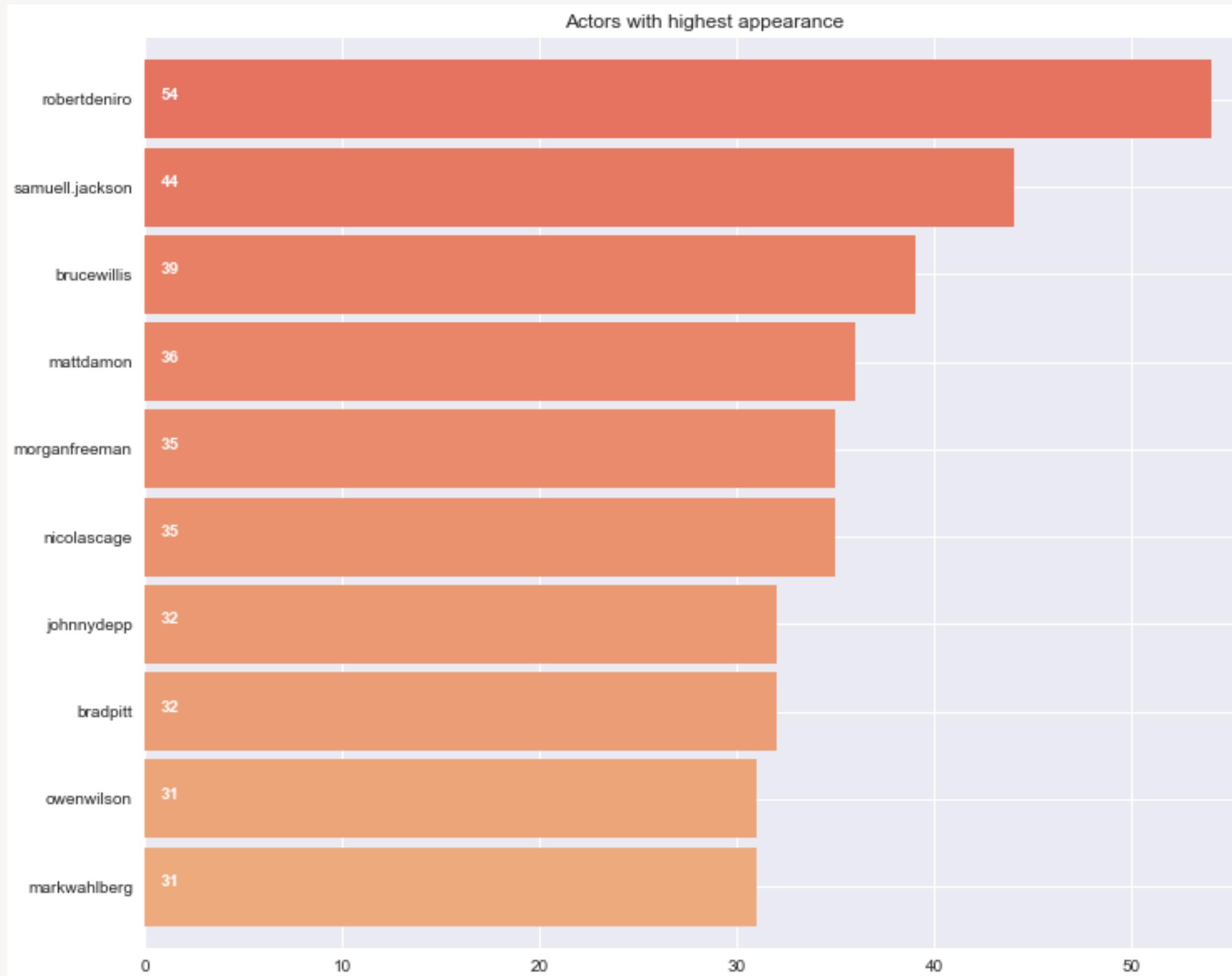
Woody Allen: 21



Clint Eastwood: 20



Top 10 Actors (in terms of appearance in this dataset)



Robert De Niro: 54



Samuel L. Jackson: 44



Bruce Willis: 39



Movie Recommendation Model



CountVectorizer

- Transform text into vector based on the frequency (count) of each word that occurs in the entire text
- Creates a matrix:
 - Column: Unique Word
 - Row: Text Sample (Movie)

Movie	action	adventure	fantasy	crime
Avatar	1	1	1	0
Pirates of the Caribbean: At World's End	1	1	1	0
Spectre	1	1	0	1

Cosine similarity uses vectors to calculate the cosine of angles which signifies degree of similarity.

- Higher Similarity = Small angles = nearer to 1
- Lower Similarity = Bigger angles = nearer to -1

Movie: The Avengers

Model: Calculate the Cosine Similarity of all movies based on the features (Cast, Director, Keywords & Genres) and locate the most similar movies (cosine similarity score nearer to 1).



CountVectorizer



Cosine Similarity



Count Vectorizer & Cosine Similarity

Count Vectorizer - The Hunger Games:

102	The Hunger Games: Mockingjay - Part 2
183	The Hunger Games: Catching Fire
200	The Hunger Games: Mockingjay - Part 1
661	Zathura: A Space Adventure
193	After Earth
91	Independence Day: Resurgence
256	Allegiant
294	Epic
1326	The 5th Wave
46	X-Men: Days of Future Past

Count Vectorizer - The Avengers:

7	Avengers: Age of Ultron
26	Captain America: Civil War
79	Iron Man 2
85	Captain America: The Winter Soldier
169	Captain America: The First Avenger
174	The Incredible Hulk
31	Iron Man 3
33	X-Men: The Last Stand
68	Iron Man
94	Guardians of the Galaxy



Next Steps

Apply the Recommender

Recommending movies based on the movie which the user has recently watched.

Tweak the Recommender

Tweak accordingly based on categories such as recommending only movies directed by the same director, or with the similar casts.

Higher Profit

Increased time spent will attract more advertisements and users which will translate into more profit.

Thank you

Capstone Project
By April Kwong

