The Movie Lens dataset

Data cleaning (Please refer to "movies.ipynb" or the equivalent python file "str_titles.py")

1. Removing the rows with no genres which reduces the shape of data frame from (62423, 3) to (57361, 3) (Cell 4).
2. Converting the type of genres column from data frame to list for different genres of movies (Cell 5).
3. Adding rows for a title of movie according to its list of genres (Cell 5). The result of the first ten rows is like below:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure
1	1	Toy Story (1995)	Animation
2	1	Toy Story (1995)	Children
3	1	Toy Story (1995)	Comedy
4	1	Toy Story (1995)	Fantasy
5	2	Jumanji (1995)	Adventure
6	2	Jumanji (1995)	Children
7	2	Jumanji (1995)	Fantasy
8	3	Grumpier Old Men (1995)	Comedy
9	3	Grumpier Old Men (1995)	Romance

4. Grouping the titles of movies according to genres column using "group by" (Cell 6). The result is as follows:

	genres	title
0	Action	[Heat (1995), Sudden Death (1995), GoldenEye (...
1	Adventure	[Toy Story (1995), Jumanji (1995), Tom and Huc...
2	Animation	[Toy Story (1995), Balto (1995), Pocahontas (1...
3	Children	[Toy Story (1995), Jumanji (1995), Tom and Huc...
4	Comedy	[Toy Story (1995), Grumpier Old Men (1995), Wa...
5	Crime	[Heat (1995), Casino (1995), Money Train (1995...
6	Documentary	[Across the Sea of Time (1995), Nico Icon (199...
7	Drama	[Waiting to Exhale (1995), American President,...
8	Fantasy	[Toy Story (1995), Jumanji (1995), City of Los...
9	Film-Noir	[Devil in a Blue Dress (1995), Suture (1993), ...
10	Horror	[Dracula: Dead and Loving It (1995), Copycat (...
11	IMAX	[Wings of Courage (1995), Across the Sea of Ti...
12	Musical	[Pocahontas (1995), Muppet Treasure Island (19...
13	Mystery	[Copycat (1995), City of Lost Children, The (C...
14	Romance	[Grumpier Old Men (1995), Waiting to Exhale (1...
15	Sci-Fi	[Powder (1995), City of Lost Children, The (Ci...
16	Thriller	[Heat (1995), GoldenEye (1995), Money Train (1...
17	War	[Richard III (1995), Misérables, Les (1995), B...
18	Western	[Desperado (1995), Wild Bill (1995), Legends o...

5. To remove numbers, punctuations, conjunctions, auxiliary verbs, prepositions, and nonsense words using NLTK library, the function “str_per_genre” is created (Cell 7). The result is as follows:

	genres	title
0	Action	Heat Sudden Death GoldenEye Cutthroat Island M...
1	Adventure	Toy Story Huck GoldenEye Cutthroat Island City...
2	Animation	Toy Story Goofy Movie Gumby Movie Swan Princes...
3	Children	Toy Story Huck Babe Two Big Green Round Table ...
4	Comedy	Toy Story Old Men Waiting Exhale Father Bride ...
5	Crime	Heat Casino Money Train Get Copycat Shanghai T...
6	Documentary	Across Sea Time Icon Madam Catwalk Frank Wife ...
7	Drama	Waiting Exhale President Casino Sense Sensibil...
8	Fantasy	Toy Story City Lost La Mortal Round Table Cupb...
9	Film-Noir	Devil Blue Dress Suture Bitter Moon Force Evil...
10	Horror	Dead Loving Copycat Dusk Till Dawn Mary Vampir...
11	IMAX	Courage Across Sea Time Lion King Beauty Beast...
12	Musical	Treasure Island de Lion King Love Got Thirty T...
13	Mystery	Copycat City Lost La Twelve k Seven k Usual Co...
14	Romance	Old Men Waiting Exhale President Cutthroat Isl...
15	Sci-Fi	Powder City Lost La Twelve k Man Beyond Unforg...
16	Thriller	Heat GoldenEye Money Train Get Copycat Twelve ...
17	War	Rob Beyond Bacon Crimson Tide Rain Fall Walkin...
18	Western	Desperado Wild Bill Fall Quick Dead Maverick B...

The results are stored in a dictionary named “dict_titles_per_genre”.

MapReduce job

1. To use the titles of each genre as input to MapReduce job, text files called “titles_i.txt” (0 <= i <= 18) are created (Cell 11).
2. To save the outputs of MapReduce job, a text file called “top10.txt” is created.
3. The implementation of MapReduce job (counting the words and giving the top ten frequent ones) is written into a python file called “map_reduce.py”.
4. The outputs of top 10 frequent words are saved in “top10.txt” file.

Commands to be executed in terminal:

1. echo "genre_name" >> “your_path”/top10.txt
2. echo "==============================" >> “your_path”/top10.txt
3. python “your_path”\map_reduce.py “your_path”\”titles_i.txt” >> “your_path”\top10.txt