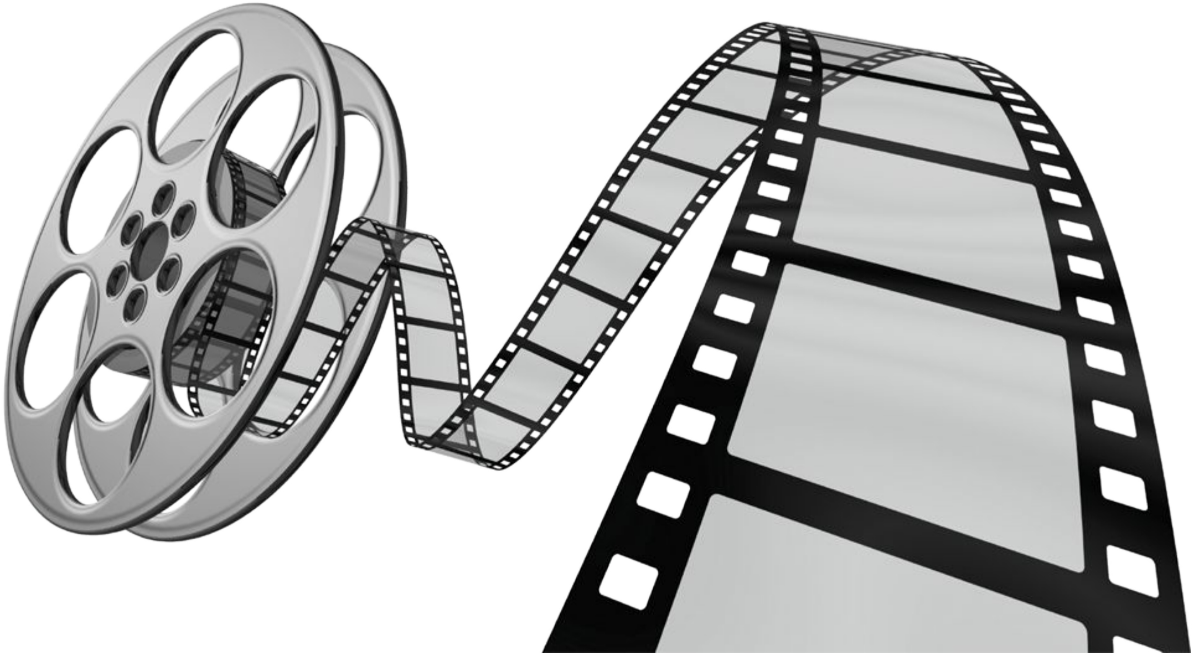


Movie Similarities, AGAIN

Team members:
Andrei Cojocaru, Diana Minzat, Nicolae Pavel

Starting ...

What are we (like the other 20 teams) trying to do:

- Recommend movies similar to another movie.
- [Recommend movies based on user rating of a movie]

What data do we use:

Database from:http://grouplens.org/datasets/movielens/

Provides 3 text files:

ratings.dat : `userid::movieid::rating::timestamp`

movies.dat : `movieID::title::genres`

users.dat : `userid::gender::age::occupation::zip-code`

Going deeper…

How are we doing it:

1. Process the files using MapReduce
2. Obtain a final file that has all movie pairs and their similarities and the number of ratings
3. Process the results, add score depending on similarity, number of ratings, and movie categories
4. Present recommandations to user

What do we use:

1. Python with mrjob package for MapReduce
2. Basic Cloudera HADOOP distribution
3. Python for final results processing and presentation

MapReduce Jobs description

- Initial data arranging (we also normalize the users ratings)
  - Map: Emit User_ID (Movie_ID, Rating)
  - Reduce: For each User_ID pair all their (Movie_ID, Rating)
- Matching Pairs
  - Map: Emit(Movie_ID, Movie_ID) (Rating, Rating) from the reducer output in Job 1 for all User_ID pairs
  - Reduce: For multiple (Movie_ID, Movie_ID) pairs calculate similarity (Pearson correlation formula)
- Similarity Results
  - Map: Emit (Movie_ID, Similarity) (Movie_ID_compared, Count) from the reducer output in Job 2
  - Reducer: Output (Movie_ID, Movie_ID_compared) (Similarity, Count)

Implementation details

Data normalization

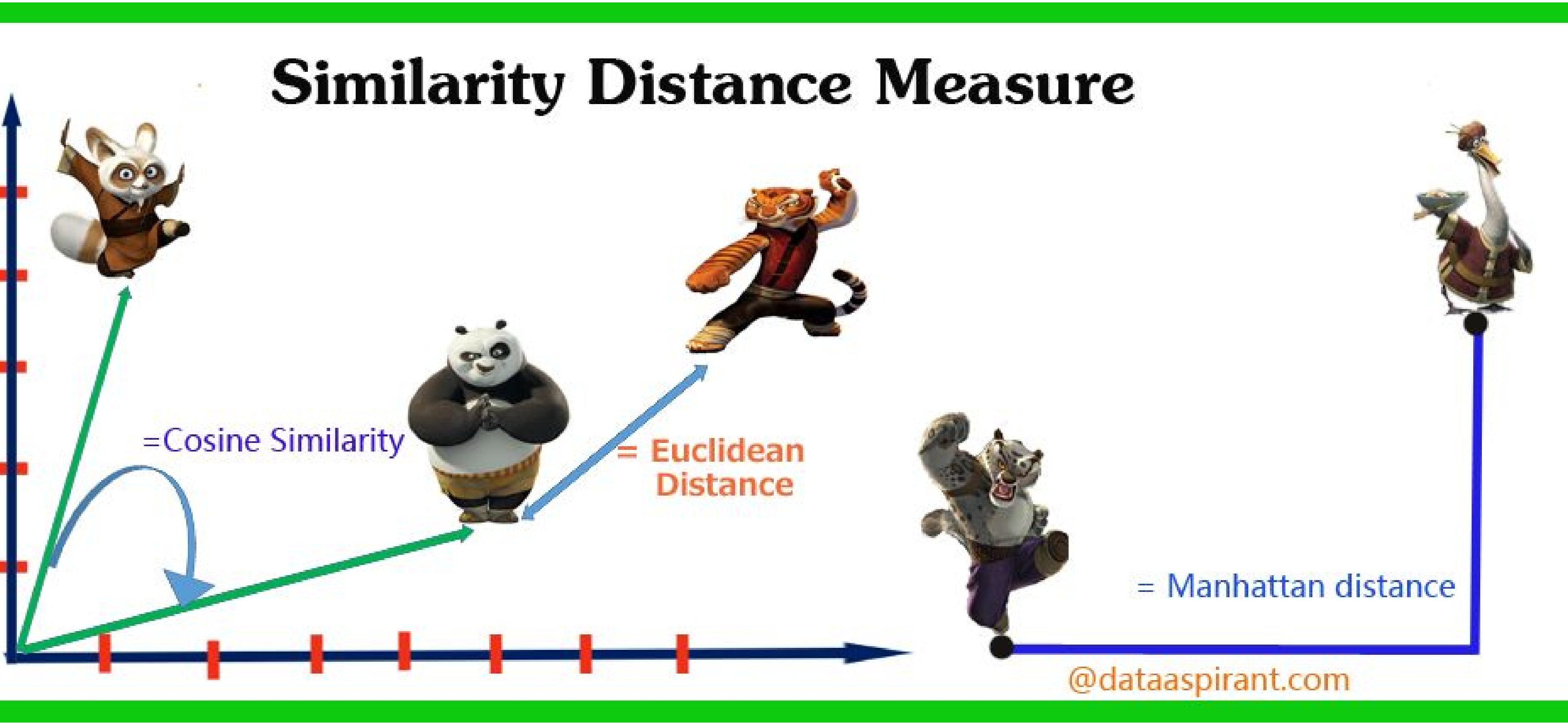
We should have normalized user ratings in range [-1,1], with a mean of 0 and a sample standard deviation of 1

$$rating_{um} = \frac{x_{um} - \bar{x}_u}{s_u}$$

New ratings are computed by subtracting the average of user's ratings and then divided by the standard deviation. This way we deal with users who give higher or lower overall ratings.

However, this proved to confuse the Pearson correlation so we removed it.

So many options



Implementation details

Similarity criteria

- Pearson product-moment correlation coefficient (*)

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Scaled correlation coefficient

$$\bar{r}_s = \frac{1}{K} \sum_{k=1}^{K} r_k,$$

- Reflective correlation coefficient

$$rr_{xy} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}.$$

Implementation details

Similarity criteria (continued):

- Jaccard similarity (*)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

- Cosine similarity (!)

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

- Manhattan similarity (*)

$$ManhattanSim(u,v) = 1 - \frac{\sum_{i=1}^{n} |u[i] - v[i]|}{n}.$$

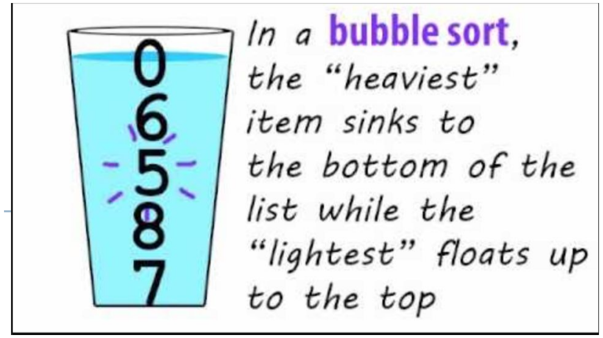
Similarity winner

Seems to be the cosine similarity, since it does recommend Alien as similar to Aliens pretty high in the results list.

Pearson would be a close second, especially on other tests (Lethal Weapon series).

SVD (not tested extensively) also at least recommends Alien.

Results, custom sorting



In a **bubble sort**, the "heaviest" item sinks to the bottom of the list while the "lightest" floats up to the top

We have 2 text based UIs for selecting and presenting results:

1. One that works directly on MapReduce output
2. One that works on a sqlite database with data imported from MapReduce output (for speed on the 1M set)

We present results sorted by similarity, number of ratings, and common categories.

We have found out experimentally that without matching movie categories the results are unrealistic (Toy Story similar to Philadelphia ?!?); number of ratings and similarity work as a confidence score.

Testing details

We generated several ratings subsets for speed of testing, a 1K “small and extra fast” set but also 100K, 250K and 900K ratings sets.

Results get more realistic with sample size, but the 1K and 100K sets were invaluable for development.

Examples (Pearson similarity)

Movie Aliens (1986) Genres: ['Action', 'Sci-Fi', 'Thriller', 'War'] is similar to:

Independence Day (ID4) (1996) Similarity 0.92 by 8 people. Score: 3

Them! (1954) Similarity 0.94 by 5 people. Score: 3

War of the Worlds, The (1953) Similarity 0.87 by 3 people. Score: 3

Westworld (1973) Similarity 1.00 by 2 people. Score: 3

Stargate (1994) Similarity 0.86 by 11 people. Score: 2

Star Trek V: The Final Frontier (1989) Similarity 0.89 by 6 people. Score: 2

Sphere (1998) Similarity 0.85 by 5 people. Score: 2

Examples (cosine similarity)

```
Movie Aliens (1986) Genres: ['Action', 'Sci-Fi', 'Thriller', 'War'] is similar to:
Soldier (1998) Similarity 1.00 by 1 people. Score: 4
Terminator, The (1984) Similarity 0.98 by 15 people. Score: 3
Alien (1979) Similarity 0.99 by 14 people. Score: 3
Matrix, The (1999) Similarity 0.98 by 14 people. Score: 3
Total Recall (1990) Similarity 0.97 by 13 people. Score: 3
Predator (1987) Similarity 0.97 by 13 people. Score: 3
```

Examples (Jaccard)

Movie Aliens (1986) Genres: ['Action', 'Sci-Fi', 'Thriller', 'War'] is similar to:

Soldier (1998) Similarity 1.00 by 1 people. Score: 4

Terminator, The (1984) Similarity 1.00 by 15 people. Score: 3

Predator (1987) Similarity 1.00 by 13 people. Score: 3

Star Wars: Episode VI - Return of the Jedi Similarity 1.00 by 9 people. Score: 3

Westworld (1973) Similarity 1.00 by 2 people. Score: 3

Men in Black (1997) Similarity 1.00 by 18 people. Score: 2

Star Trek: First Contact (1996) Similarity 1.00 by 12 people. Score: 2

Examples (Manhattan)

Movie Aliens (1986) Genres: ['Action', 'Sci-Fi', 'Thriller', 'War'] is similar to:

Soldier (1998) Similarity 1.00 by 1 people. Score: 4

Them! (1954) Similarity 0.80 by 5 people. Score: 3

Westworld (1973) Similarity 1.00 by 2 people. Score: 3

Star Trek V: The Final Frontier (1989) Similarity 0.92 by 6 people. Score: 2

Conspiracy Theory (1997) Similarity 0.80 by 5 people. Score: 2

Payback (1999) Similarity 0.80 by 5 people. Score: 2

Last of the Mohicans, The (1992) Similarity 0.83 by 3 people. Score: 2

Guns of Navarone, The (1961) Similarity 0.83 by 3 people. Score: 2

Examples (SVD)

Or “that’s why I went to bed at 5 am”.

Movie Aliens (1986) Genres: ['Action', 'Sci-Fi', 'Thriller', 'War'] is similar to:

Them! (1954) Similarity 0.38 by 1 people. Score: 3

Terminator, The (1984) Similarity 0.24 by 1 people. Score: 3

Alien (1979) Similarity 0.23 by 1 people. Score: 3

Guns of Navarone, The (1961) Similarity 0.27 by 1 people. Score: 2

What Ever Happened to Baby Jane? (1962) Similarity 0.48 by 1 people. Score: 1

More Examples (Pearson)

Movie Lethal Weapon 4 (1998) Genres: ['Action', 'Comedy', 'Crime', 'Drama'] is similar to:

Lethal Weapon 2 (1989) Similarity 0.81 by 6 people. Score: 4

Lethal Weapon 3 (1992) Similarity 0.98 by 3 people. Score: 4

Breakfast Club, The (1985) Similarity 1.00 by 4 people. Score: 2

Bodyguard, The (1992) Similarity 1.00 by 3 people. Score: 2

More examples (cosine)

```
Movie Lethal Weapon 4 (1998) Genres: ['Action', 'Comedy', 'Crime', 'Drama'] is similar to:
Lethal Weapon 2 (1989) Similarity 0.95 by 6 people. Score: 4
Lethal Weapon (1987) Similarity 0.99 by 4 people. Score: 4
Lethal Weapon 3 (1992) Similarity 0.99 by 3 people. Score: 4
Get Shorty (1995) Similarity 0.93 by 6 people. Score: 3
Batman (1989) Similarity 0.97 by 5 people. Score: 3
Midnight Run (1988) Similarity 0.89 by 4 people. Score: 3
Untouchables, The (1987) Similarity 0.94 by 3 people. Score: 3
French Connection, The (1971) Similarity 0.85 by 3 people. Score: 3
```