Homework #1

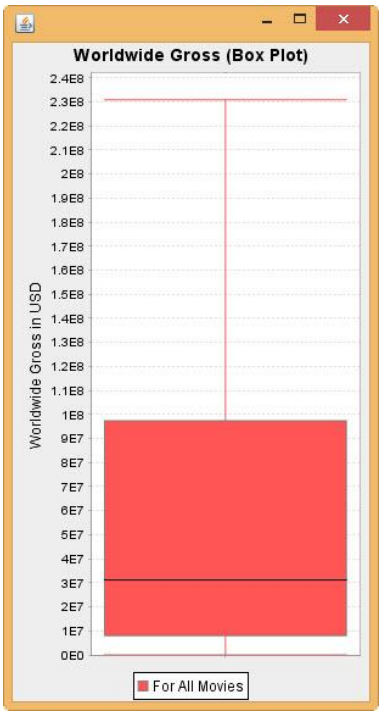
CS 5890, Fall 2015

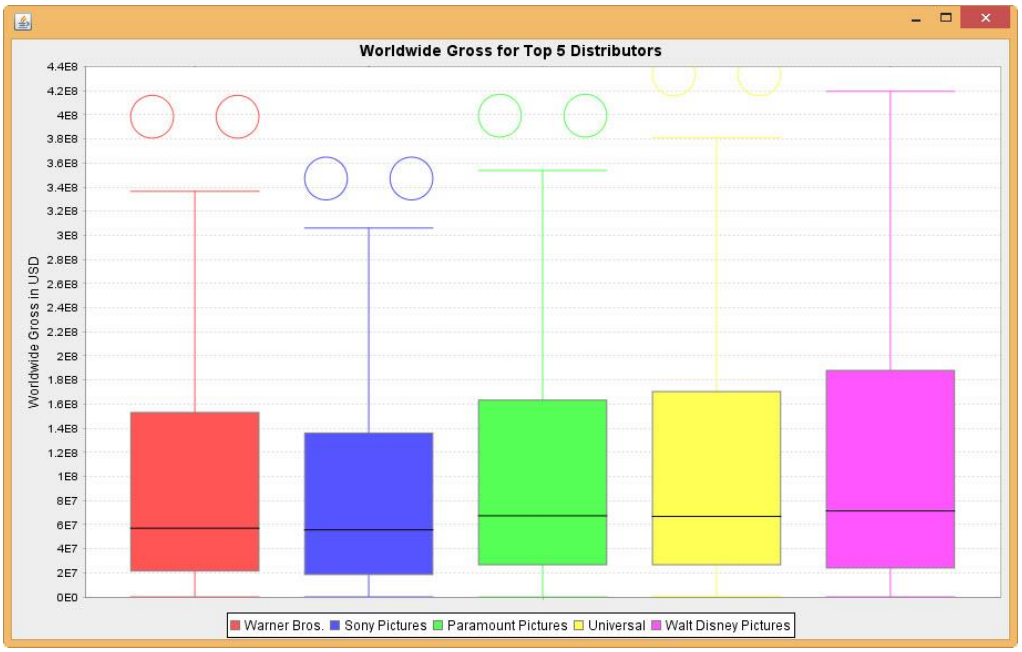
For this assignment I have carried out following stuffs:

- For this assignment I have used ***Java*** on eclipse IDE and ***JFreeChart*** and ***Tableau*** for data visualization.
- First, I have parsed the csv file containing the dataset using custom parser into an ArrayList of MovieInfo objects. MovieInfo class contains all the attributes of each movie title.
- For handling missing and invalid attributes, I have put checks in my java parsing code. For example, if an integer attribute is missing, I have replaced it with -1(flag), indicating that it should be considered as missing attribute. Similarly for missing "String" or "Nominal" attribute I have replaced it with "NA".
- For the nominal attributes, I have assigned unique integer values to all the different attribute values.
- Lastly, I have normalized all the attributes to between 0 and 1.

1. ***World-wide Gross Analysis:*** *First, we want to gain an understanding of the worldwide grosses of movies, and how these may vary across distributors.*

   (a) *Using the three main measures of central tendency (mean, median and mode), analyze the worldwide grosses for all movies, as well as for individual distributors (say, for the top-5 distributors). You should plot box-plots for all movies, as well as for the individual distributors.*





Answer 1: The left box plot shows the distribution of all the movies contained in the CSV and the one on the right shows distribution for top 5 distributors, which are 1.Warner Bros., 2.Sony Pictures, 3.Paramount, 4.Universal and 5. Walt Disney (in increasing order). I have made following conclusion after analyzing the charts:

- First of all, I was surprised by the box plot's maximum value (Upper Whisker), 2.3E8. It is nowhere close to the actual value of the top grossing movie, which 2.7E9. After going through online texts about box plots, I found out about a new term called 'outliners'. Outliners are values that are much bigger or smaller than the rest of the data. These are represented by a dot at either end of the plot. In order to be an outlier, the data value must be: 1) larger than Q3 by at least 1.5 times the interquartile range (IQR), or 2) smaller than Q1 by at least 1.5 times the IQR. That's the thumb rule. So, in our dataset all the values above 2.3E8 has been considered to be outliners. And have been shown above top whiskers as large circles.
- Because of these 'outliners' the box plots aren't very reliable, and don't provide accurate information.
- On observing the box plot on the left for all the movies, we can see that IQR is more towards the lower values in the range of 1E7 and 1E8, which indicates that there are more movies with less worldwide gross. The median is even lower, 3E7. The whole thing concludes that there are more low grossing movies in comparison to higher grossing ones.
- The mean, median and mode of the "All movies" are: **Mean = 85343400, Median = 31168926, Mode = 0**.
- On observing the right box plot for top 5 distributors, we can see that IQR for Walt Disney Pictures is more spread-out than other 4 distributors, which indicates that Walt Disney Pictures has distributed movies which have quite varying world gross. Also maximum value (top whisker) is quite high, indicating that it has Disney has produced some high grossing movies than other distributors. Similarly, IQR, mean and top whisker for Sony pictures are smaller than others, indicating that movies distributed by it are low grossing movies.

(b) *Now, remove outliers by dropping all grosses that are "too extreme." Be sure to quantify your definition of "too extreme" and explain how you arrived at that definition. Compare the mean, median, and mode without outliers. What do you observe?*

Answer 1 (b):

"Too Extreme" Values: The "Inter Quartile Range (IQR)" is the length of the box in box-and-whisker plot. An outlier is any value that lies more than one and a half times the length of the box from either end of the box. That is, if a data point is below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$, it is viewed as being too far from the central values to be reasonable, and are considered as "Extreme Values". In our box plot, the upper whisker is at 234800000. So any value above 234800000 is considered as "Too Extreme". Also, the bottom whisker is at 0, so any values below this, i.e. negative values are considered at "Too Extreme" as well. After ignoring all the values, we get:

Mean: 47133715
Median: 24911670
Mode: 0

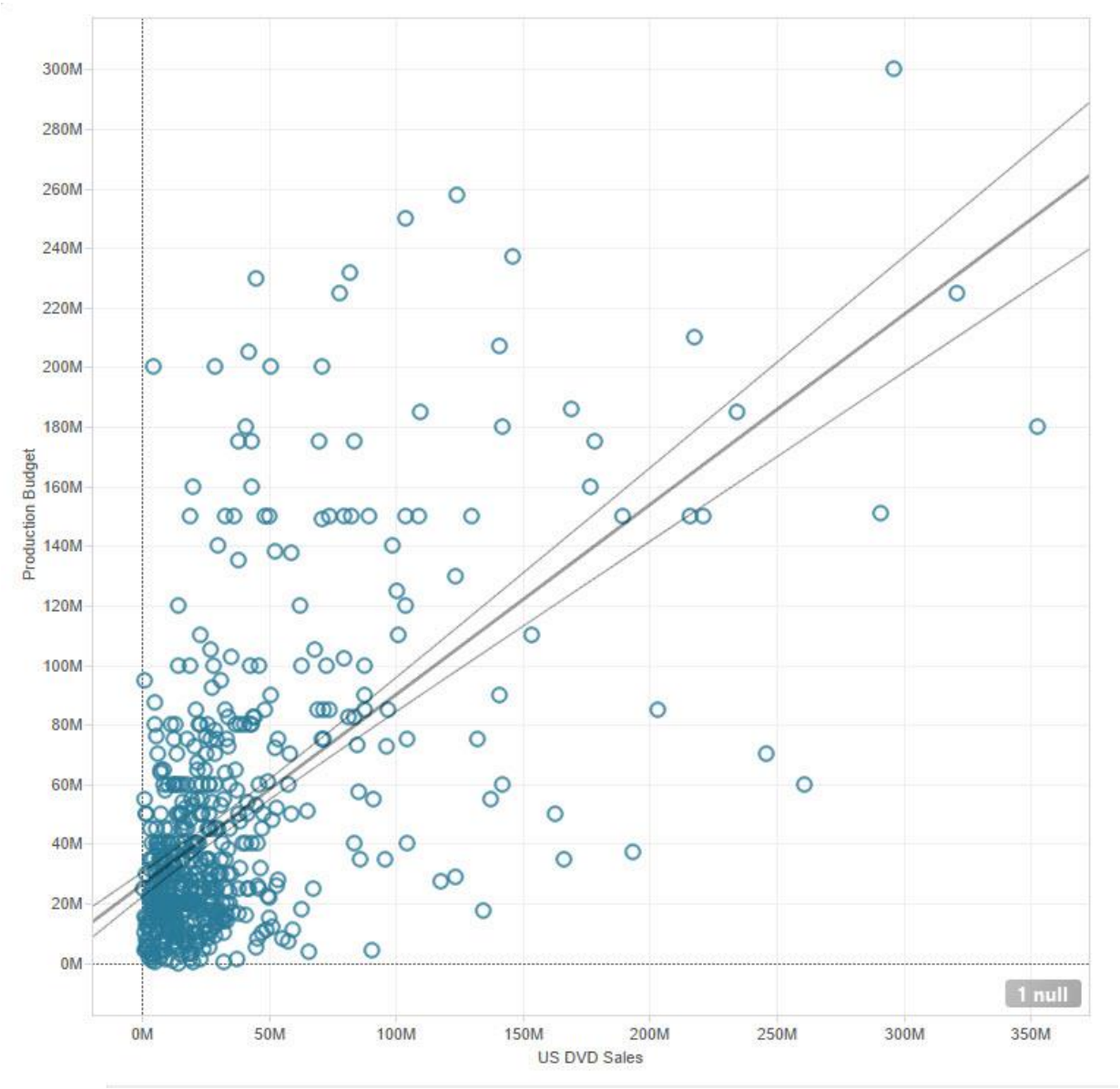
We can observe that mean comes down significantly. Median becomes less too. But mode remains the same.

--------- x ----------

2. **Revenue Correlation***: Does a relationship hold between the US DVD sales for a movie and it's Production Budget?*

(a) *Plot this relationship using a scatter plot, and report the correlation (using Person's correlation coefficient).*

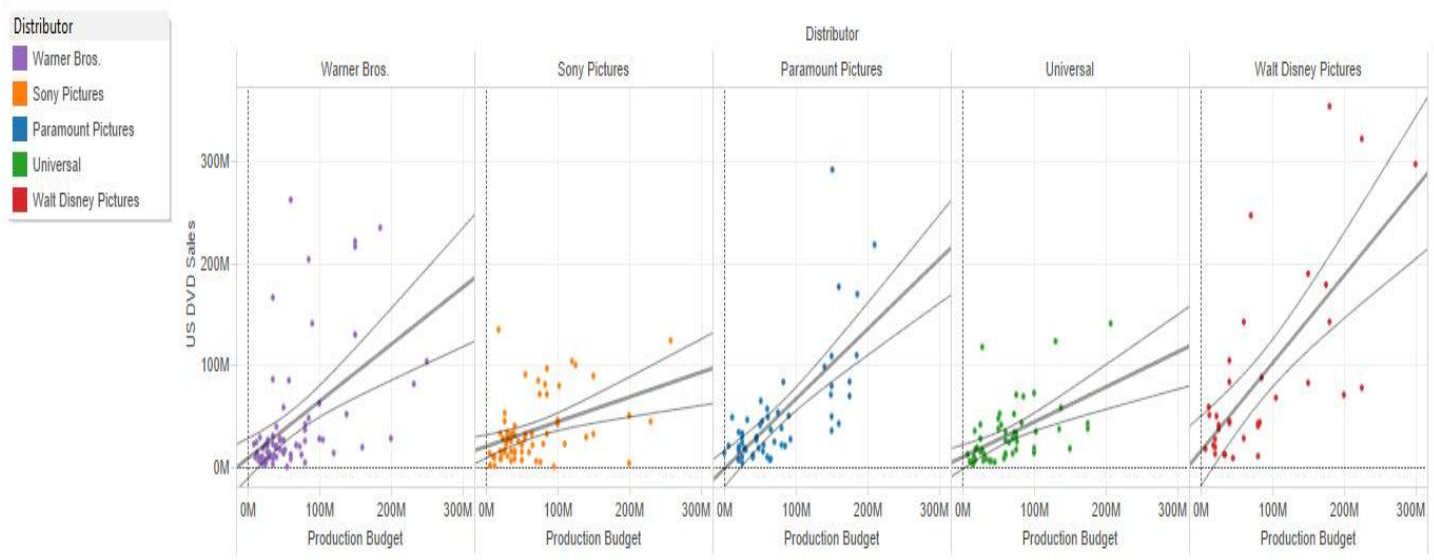
Answer 2 (a): Below is the scatter plot between US DVD sales for a movie and its production budget.



- On using the Pearson's Correlation formula the value comes out to be: **0.591,** which shows that there's a strong positive relationship between DVD Sales and Production Budget.
- We can infer from the graph that most of the movies have been low budget movies, as the points are clustered towards the origin.

(b) *Now find the top 5 distributors based upon the number of movies and perform the same analysis for these 5 distributors. Based upon these plots and Pearson correlation values what can you conclude?*

Answer 2 (b): Below is the scatter plot for top 5 distributers based upon the number of movies:



The Pearson's coefficient for top 5 distributors are as follows:

- Warner Bros.: 0.49
- Sony Pictures: 0.415
- Paramount Pictures: 0.71
- Universal: 0.534
- Walt Disney: 0.707

Following points can be inferred from the scatter graph and Pearson's coefficient values:

- **Sony Pictures**: The Pearson's coefficient of Sony Pictures (0.415) is the lowest amongst other distributors. This means that, even though Sony Pictures produces high budget movie, its DVD sales does not increase significantly. Also, the dots are clustered towards the origin point, which means that Sony distributes lower budget movies.
- **Universal** & **Warner Bros**: In both these graphs, the dots are clustered towards the original, so both of these companies distribute lower budget movies. But with warner bros, there are few low budget movies that has given some very good DVD sales, as indicated by dots towards high y-axis.
- **Paramount Pictures:** Here we can see two clusters, one big cluster close to the origin and other smaller one in the center. Bigger cluster indicates that paramount pictures produces lower budget movies, but the DVD sales is not good for these. The other sparser cluster in vertically spread and is tightly bound horizontally, meaning Paramount distributes some high budget movies and they do very well in sales. The Pearson's coefficient is 0.71, meaning that there is a high positive correlation.
- **Walt Disney Pictures:** Here the dots are spread all over the space, there is no clustering. This indicates that Disney distributes movies of wide range of budgets. And most of them have good DVD sales. There are few high budget movies, indicated by dots towards the top right corners, which have done exceptionally well for the DVD sales. Even the Pearson's coefficient is quite high (0.707).

--------- x ----------

3. Movie Similarities: *Using two distance metrics (Euclidean and Manhattan) and one similarity function (Cosine), find the 5 movies closest to following movies:*

- *The Matrix*
- *The Godfather*
- *The Shawshank Redemption*
- *Toy Story*
- *The Lord of the Rings: The Fellowship of the Ring*

Answer 3: The similar movies using the dimensions provided in the question are as follows (The similarity of each 5 movies is in descending order, i.e. the leftmost movie is most similar, and so on for right movies):

1) **The Matrix**:
   i. General Movie Info:
      - *Euclidean Distance*: The Matrix Reloaded, The Negotiator, The Rock, The Matrix Revolutions, Blood Diamond
      - *Manhattan Distance*: The Matrix Reloaded, The Negotiator, The Rock, The Matrix Revolutions, Blood Diamond
      - *Cosine Similarity*: The Rock, The Negotiator, The Matrix Reloaded, Blood Diamond, The Matrix Revolutions
   ii. Rating Only:
      - *Euclidean Distance*: Fight Club, Pulp Fiction, The Godfather, American Beauty, The Dark Knight
      - *Manhattan Distance*: Fight Club, Pulp Fiction, American Beauty, The Usual Suspects, The Dark Knight
      - *Cosine Similarity*: Pulp Fiction, Fight Club, The Godfather, The Dark Knight, Gladiator
   iii. Revenue Only:
      - *Euclidean Distance*: Tarzan, Pretty Woman, Mission: Impossible, Ocean's Eleven, Die Another Day
      - *Manhattan Distance*: Tarzan, Pretty Woman, Mission: Impossible, Ocean's Eleven, Rain Man
      - *Cosine Similarity*: Mars Attacks!, Die Another Day, The Mummy, Iris, The League of Extraordinary Gentlemen
   iv. All Three Types Above:
      - *Euclidean Distance*: Gladiator, Terminator 2: Judgment Day, Back to the Future, Inglourious Basterds, Pirates of the Caribbean: The Curse of the Black Pearl
      - *Manhattan Distance*: Gladiator, Terminator 2: Judgment Day, Inglourious Basterds, Back to the Future, The Rock
      - *Cosine Similarity*: Gladiator, Terminator 2: Judgment Day, Back to the Future, Inglourious Basterds, Pirates of the Caribbean: The Curse of the Black Pearl

2) **The Godfather**:
   i. General Movie Info:
      - *Euclidean Distance*: The Trouble With Harry, The Godfather: Part II, Popeye, Out of the Dark, Paa
      - *Manhattan Distance*: Trouble With Harry, The Godfather: Part II, Popeye, Out of the Dark, Paa
      - *Cosine Similarity*: The Trouble With Harry, The Godfather: Part II, Popeye, Out of the Dark, Paa

ii. Rating Only:
- *Euclidean Distance*: Pulp Fiction, The Dark Knight, The Matrix, Fight Club, The Shawshank Redemption
- *Manhattan Distance*: Pulp Fiction, The Dark Knight, The Matrix, Fight Club, Schindler's List
- *Cosine Similarity*: Pulp Fiction, The Matrix, Fight Club, The Dark Knight, American Beauty

iii. Revenue Only:
- *Euclidean Distance*: The Incredible Hulk, Bad Boys II, Lara Croft: Tomb Raider, The Silence of the Lambs, American Gangster
- *Manhattan Distance*: The Incredible Hulk, Bad Boys II, Lara Croft: Tomb Raider, The Silence of the Lambs, American Gangster
- *Cosine Similarity*: The Ice Storm, It's Complicated, Miss Congeniality, Iron Man 2, The Karate Kid

iv. All Three Types Above:
- *Euclidean Distance*: Reservoir Dogs, The Godfather: Part II, Pulp Fiction, Se7en, District 9
- *Manhattan Distance*: The Godfather: Part II, Reservoir Dogs, Pulp Fiction, Singin' in the Rain, The Great Escape
- *Cosine Similarity*: Se7en, Reservoir Dogs, Pulp Fiction, The Usual Suspects, American History X

3) **The Shawshank Redemption**:

i. General Movie Info:
- *Euclidean Distance*: Being Julia, Legends of the Fall, Mary Reilly, The End of the Affair, Auto Focus
- *Manhattan Distance*: Being Julia, Legends of the Fall, Mary Reilly, The End of the Affair, Auto Focus
- *Cosine Similarity*: The End of the Affair, Mary Reilly, Legends of the Fall, eing Julia, The Merchant of Venice

ii. Rating Only:
- *Euclidean Distance*: The Dark Knight, Pulp Fiction, The Godfather, The Matrix, Fight Club
- *Manhattan Distance*: The Dark Knight, Pulp Fiction, The Godfather, The Matrix, Fight Club
- *Cosine Similarity*: The Dark Knight, Fight Club, Pulp Fiction, The Matrix, The Godfather

iii. Revenue Only:
- *Euclidean Distance*: Friday, The Brothers, Love and Basketball, Poetic Justice, Dune
- *Manhattan Distance*: Girl, Interrupted, Friday, The Brothers, Brown Sugar, The New Guy
- *Cosine Similarity*: Turbulence, The Wood, The Wash, Tales from the Hood, Superman II

iv. All Three Types Above:
- *Euclidean Distance*: Goodfellas, Memento, Schindler's List, Forrest Gump, The Silence of the Lambs
- *Manhattan Distance*: Goodfellas, Memento, Casino, The Shining, Trainspotting
- *Cosine Similarity*: Memento, Goodfellas, Schindler's List, Forrest Gump, Braveheart

4) **Toy Story**:

i. General Movie Info:
- *Euclidean Distance*: Gake no ue no Ponyo, The Lion King, The Wild, Valiant, Toy Story 3
- *Manhattan Distance*: Gake no ue no Ponyo, The Lion King, The Wild, Valiant, Toy Story 3
- *Cosine Similarity*: Valiant, The Wild, The Lion King, Gake no ue no Ponyo, Toy Story 3

ii. Rating Only:

- *Euclidean Distance*: Jaws, Good Will Hunting, Finding Nemo, The Incredibles, L.A. Confidential
- *Manhattan Distance*: Jaws, Good Will Hunting, Finding Nemo, The Incredibles, L.A. Confidential
- *Cosine Similarity*: Spider-Man 2, Good Will Hunting, Minority Report, Juno, The Incredibles

iii. Revenue Only:

- *Euclidean Distance*: The Fugitive, What Women Want, Gone with the Wind, Batman Forever, Jurassic Park 3
- *Manhattan Distance*: The Fugitive, What Women Want, Gone with the Wind, Batman Forever, Jurassic Park 3
- *Cosine Similarity*: The 13th Warrior, Species, Enemy at the Gates, Dark City, Blade 2

iv. All Three Types Above:

- *Euclidean Distance:* Toy Story 2, A Bug's Life, The Lion King, Finding Nemo, Monsters, Inc.
- *Manhattan Distance*: Toy Story 2, A Bug's Life, The Lion King, Finding Nemo, Monsters, Inc.
- *Cosine Similarity*: Toy Story 2, A Bug's Life, The Lion King, Monsters, Inc., Finding Nemo

5) **The Lord of the Rings: The Fellowship of the Ring**:

i. General Movie Info:

- *Euclidean Distance*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, K-19: The Widowmaker, The Man in the Iron Mask, Die Another Day
- *Manhattan Distance*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, The Man in the Iron Mask, Stardust, K-19: The Widowmaker
- *Cosine Similarity*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, Nochnoy dozor, The Man in the Iron Mask, Die Another Day

ii. Rating Only:

- *Euclidean Distance*: The Lord of the Rings: The Return of the King, The Lord of the Rings: The Two Towers, Se7en, Memento, Forrest Gump
- *Manhattan Distance*: The Lord of the Rings: The Return of the King, The Lord of the Rings: The Two Towers, Se7en, Memento, Fight Club
- *Cosine Similarity*: The Lord of the Rings: The Return of the King, The Lord of the Rings: The Two Towers, Se7en, Memento, Batman & Robin

iii. Revenue Only:

- *Euclidean Distance* Star Wars Ep. VI: Independence Day, Finding Nemo, The Lion King, Harry Potter and the Sorcerer's Stone, The Lord of the Rings: The Two Towers
- *Manhattan Distance*: Independence Day, Indiana Jones and the Kingdom of the Crystal Skull, Finding Nemo, Shrek the Third, Spider-Man 3
- *Cosine Similarity*: The Saint, Godzilla, Octopussy, Fly Me To the Moon, Kids

iv. All Three Types Above:

- *Euclidean Distance*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, Forrest Gump, Sin City, Batman Begins
- *Manhattan Distance*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, Forrest Gump, Harry Potter and the Sorcerer's Stone, Iron Man 2
- *Cosine Similarity*: The Lord of the Rings: The Two Towers, The Lord of the Rings: The Return of the King, Forrest Gump, Sin City, Batman Begins

To measure the distance between the movies I carried out following steps:

- First of all, most of the attributes in the movie are quantitative and few are nominal.

- Some quantitative attributes are too large (in million, like Worldwide Gross) and others are comparatively too small (<10, like IMDB rating). If these contrasting attributes are passed to the distance vectors, ***the larger attribute becomes dominant*** and the distance moves closer to the vector with larger attribute. For example, when calculating the distance between movies using the "Rating Only" attributes (Rotten Tomatoes, Rating, IMDB Rating and IMDB votes), ***IMDB votes are in order of 10000s or even 100000s and IMDB ratings between 1 to 10. So, here the IMDB votes dominates the IMDB rating, and we get movies which are similar only in IMDB votes ONLY, not all 3 attributes***.

- To overcome the above problem, I have ***normalized all the quantitative attributes to between 0 and 1***. So, after the normalization, all attributes have same value and none of them dominate in the distance calculation.

- There some nominal attributes too, like MPAA rating, Distributor, etc. To make these nominal values compatible with the distance calculation, I have assigned unique consecutive integer values to the each attribute set**. *For example, "Major Genre" attribute has values like, action, adventure, etc. So I have assigned 1 to all the "action" values, 2 to adventure, 3 to horror and so on. And after this I have normalized the values between 0 and 1.***

So, which similarity pair performs the best?

- **Short answer*: Depends on the value of attributes and their relation amongst them.*** For some attribute sets Manhattan and Euclidean perform better and for other attribute sets Cosine performs better.

- ***If the values of dimension of attribute vectors are unrelated and these values are far apart then Cosine distance measurement performs little better than Euclidean and Manhattan***. For example, while calculating the distance measurement using "*Ratings only: Rotten Tomatoes Rating, IMDB Rating, IMDB Votes*", The Values for 1st movie, i.e., "The Matrix", whose Rotten Tomatoes Rating is: 86, IMDB Rating is: 8.7 and IMDB Votes is: 380934. We can see that the difference in these values in quite large, especially between the rating and votes. All the 3 distance measurement methods (Euclidean, Manhattan and Cosine) gave out the same result, i.e. "Pulp Fiction", whose RT rating, IMDB rating and IMDB votes are 81, 8.8 and 382470 respectively. So, we conclude that all 3 distance measurement technique provide the same result when the attributes are far apart.

- ***Cosine similarity terribly fails when the value of vector dimensions are same***. For example, For example, while calculating the distance measurement using "*Revenue only: US Gross, Worldwide Gross*" for the movie The Shawshank Redemption, whose US Gross is 28241469 and Worldwide Gross is also the same 28241469. When calculating the distance cosine, the output is surprising. We get 100s of movies whose US Gross and Worldwide Gross are same, like "Turbulence", whose US Gross: 11532774 and Worldwide Gross: 11532774. Also "Warlock", whose US Gross: 8824553 and Worldwide Gross:

8824553. ***So we can see that, Cosine calculates the distance based on, if the dimension of vector are same or not, NOT on the closeness in the values of each dimensions.*** This is not what we may want.

- More formally, ***Cosine measures the similarity of vectors with respect to the origin, while the Euclidean measures the distance between particular points of interest along the vector.***

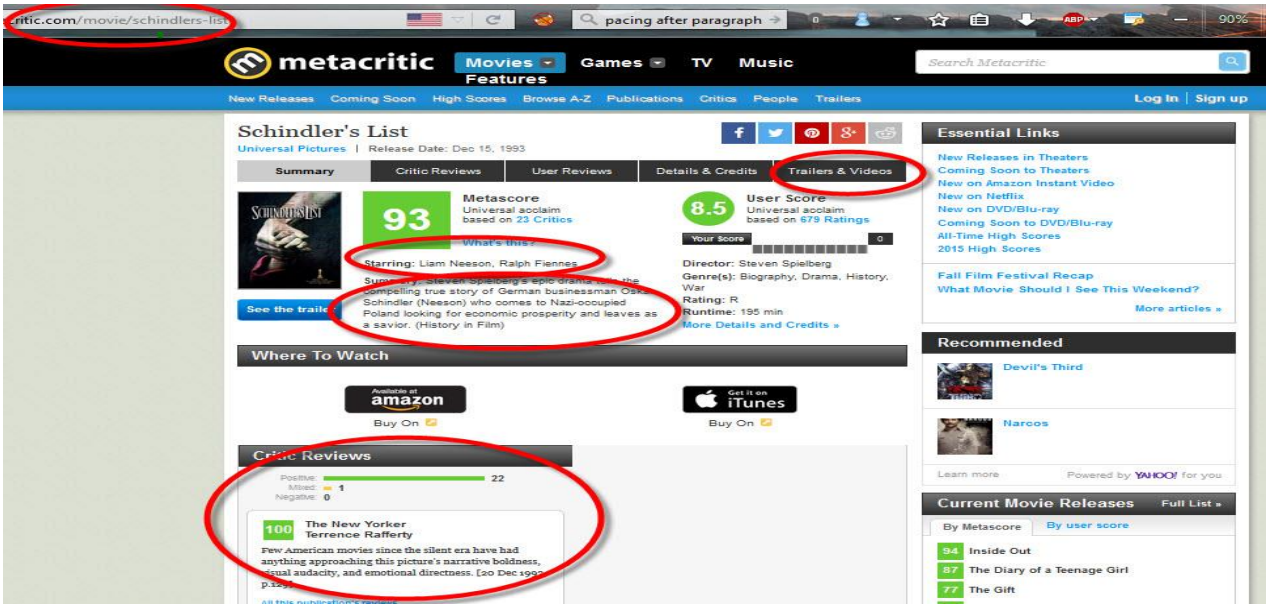
- The Euclidean/Manhattan and Cosine both provide us with a different aspect of similarity between two objects. It is up to us to either use them individually or in combination depending upon our application needs.

--------- X ----------

4. ***Augmenting the Movie Data:*** *Finally, you should identify a new dataset that could be potentially integrated with the existing movie data. You should provide a link and description of the new dataset, and identify any issues you might have in integrating the new data. (Note, you do not need to actually integrate the two datasets; for this homework, a description is sufficient).*

Answer 4:

For augmenting dataset we can use metacritic website to get additional data for the movies in our existing dataset. Link: www.metacritic.com/movie/schindlers-list.



Description about our 'to be integrated' dataset, i.e. data from metacritic:

- Metacritic.com is simple HTML website with few java scripts imbedded here and there.
- For getting the URL of any movie, we just need to append the name of the movie at the end of www.metacritic.com/movie/ (spaces in the movie title are replaced by "–"). For example, the URL of the movie Schindler's List would be www.metacritic.com/movie/schindlers-list.
- We can use data extraction tool such as Web-Harvest or jsoup to extract data from metacritic website and store it in any database or CSV file.

- From metacritic.com, we can get additional attributes for the movies in our existing dataset like, Cast, Language, Summary, Metascore rating, link to the trailer, reviews from top newspapers/magazines, etc.
- I would like to extend our existing dataset by adding following attribute fields from metacritic: Cast1, Cast2, Cast3, Summary, Review1, Review2, Review3 and TrailerURL.
- For doing this, we first need to harvest the attributes; top 3 cast members, summary, top 3 reviews and url for the trailer for each movie; and then save these fields to a file, say metacritic.csv.
- After getting the new dataset, i.e. metacritic.csv from metacritic.com, we write a function that loops through the metacritic.csv file and appends the new attributes to each movie in the movies.csv by matching the movie titles.
- Now we have a new updated movie.csv dataset containing the additional attributes.

Issues that might be encountered while integrating the data from metacritic:

- There are few movies at metacritic where there is no values for some attributes, like some movies have only 1 or 2 reviews, so we'll have to manage how we handle the value for the 3rd review field. We could either leave it empty or give it a default value like, N/A.
- In metacritic website we can see that there are more than 3 cast members for every movie. In our augmented dataset we are only using top 3 cast members, which is easy to integrate in our csv file by creating 3 new comma separated values. But what if we want to more than 3 cast members, say 5 or may be 25? This is not be feasible for csv because we'll have to create 5 or 25 new comma separated attributes. A better solution will be to use a relational database for such large values.
- Again, the values of Cast1, Cast2, Cast3, Summary, Review1, Review2, Review3 and TrailerURL, etc are nominal. So if we want to analyze these we'll need to normalize them.
- Though this is not related to data integration, but it's quite an important point. Over time, metacritic could change the structure of their website, so our harvesting function may become invalid. To overcome this, we'll need to regularly update our harvesting API, so that it continues to work.

-------- X ---------