HomeWork Assignment 1

Problem 1:

(A.) Encoding in term frequency-inverse document provides the knowledge that how much information we gain about any particular document given a particular term. i.e. Let's say we are searching for any document with a particular term than what are the chances of finding that document. Encoding equation:

$$x_{ij} = \frac{m_{ij}}{m_i} \cdot log \frac{n}{n_j}$$

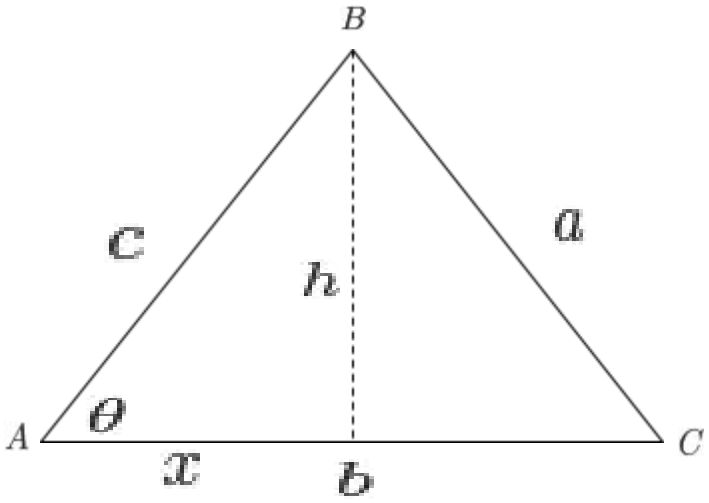
Here, $\frac{m_{ij}}{mi}$ gives the probability of occurrence of a word in the i^{th} document. So, if this quantity will be more, we can be more sure about the document given the term. Second part, $log \frac{n}{n_j}$, tells us that this term occurs in how many documents. So, the more number of documents in which this term will occur, less we can be sure about this document as the log value of this term will decrease. In short, to gain more information about the document m_{ij} value should be high and n_j value should be less.

(B.) Disadvantages of this encoding will be in the cases where a certain term appears more prominently in one document but occurs less frequently in large number of documents. Let's say, a term appears in a document, where total number of terms are 110, 100 times and in other 99 documents 1 time per document. So, total number of documents are 100 and this particular term exist in all the documents. So, the log value of second part will be **0**. So, even x_{ij} will be zero. But certainly that is not the case here. Given this particular term, we can be more sure about this document contrary to the notion suggested by this encoding. On the other hand, $\frac{m_{ij}}{m_i}$ and m_{ij} will give better information about the document.

(C.) If this term occurs in every document, the second part will be $log \frac{n}{n}$. And so, this value will be zero. It means that we would gain no information about the particular document.. On the other hand, If this term appears in just one document then log value will be very high and we can be sure about the document. So, in second case information gain would be very high.

I did the high level discussion for question no. 2 and 3 with Prateek Bhat and Ritesh Agarwal. Although, I wrote solution independent of their solution.

Problem 2:



Now, Let's assume $\measuredangle BAC = \theta$. By pythagoras theorem, we can write:

$$a^2 = h^2 + (b - x)^2$$

we can write $h = c \sin\theta$, so:

$$a^2 = c^2 sin^2\theta + (b - x)^2$$

we can also write $x = c\cos\theta$, so the equation will be:

$$a^2 = c^2 sin^2\theta + (b - c\cos\theta)^2$$

Now, expanding the square term would give:

$$a^2 = c^2 sin^2\theta + b^2 + c^2\cos^2\theta - 2bc\cos\theta$$

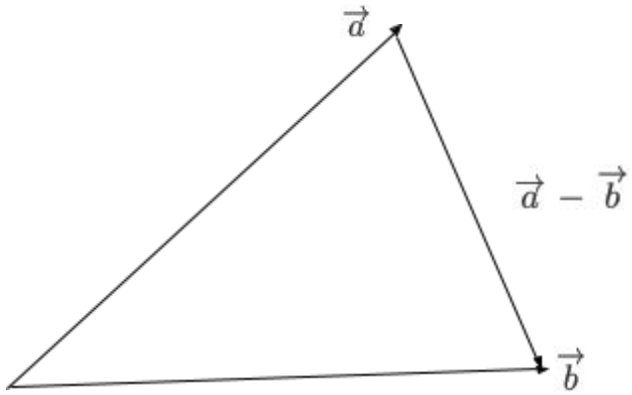
since,

$$sin^2\theta + \cos^2\theta = 1$$

so, we can write:

$$a^2 = c^2 + b^2 - 2bc\cos\theta$$

Now, for two vectors:



so, the difference between two vectors $\vec{a}$ and $\vec{b}$ will be third vector $\vec{a} - \vec{b}$.

we can write from the equation derived earlier that:

$$|\vec{a} - \vec{b}|^2 = |\vec{a}|^2 + |\vec{b}|^2 - 2|\vec{a}||\vec{b}|cos\theta$$

We can also write:

$$|\vec{a} - \vec{b}|^2 = (\vec{a} - \vec{b}) \cdot (\vec{a} - \vec{b})$$

$$= \vec{a} \cdot \vec{a} + \vec{b} \cdot \vec{b} - 2 \cdot \vec{a} \cdot \vec{b}$$

$$= |\vec{a}|^2 + |\vec{b}|^2 - 2 \cdot \vec{a} \cdot \vec{b}$$

Now, by comparing both the equations we can write:

$$2 \cdot \vec{a} \cdot \vec{b} = 2|\vec{a}||\vec{b}| \cos\theta$$

So, by cancelling each side we can write:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

We know that for k-dimensional vector, we can write dot product as:

$$\vec{a} \cdot \vec{b} = a^T . b$$

so,

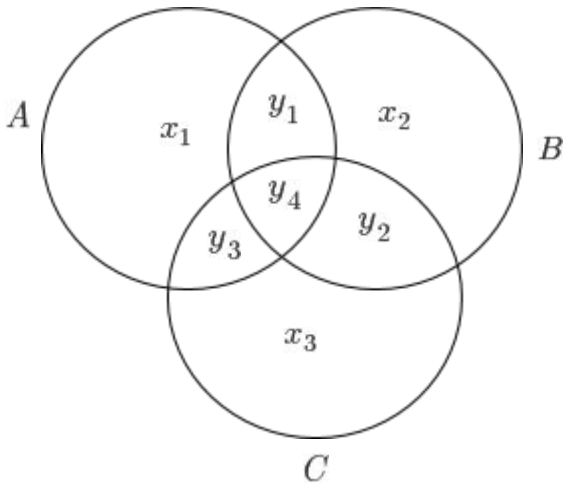
$$cos\,\theta = \frac{a^T b}{|\vec{a}|.|\vec{b}|}$$

.

Hence, proved.

Problem 3:

For this problem, we have 3 sets:



Let's assume that: $|A| = x_1$ $|B| = x_2$ $|C| = x_3$

length of exclusive intersection between set A and B $= y_1$

length of exclusive intersection between set B and C $= y_2$

length of exclusive intersection between set A and C $= y_3$

And , length of common intersection between all three sets be: y_4

So, $|A \cap B| = y_1 + y_4$ and so on..

(A.) $d_1(A,B) = |A - B| + |B - A|$

So, we need to prove that:

$$|A - B| + |B - A| + |B - C| + |C - B| \geq |A - C| + |C - A|$$

So, As per the notation:

$$x_1 - (y_1 + y_4) + x_2 - (y_1 + y_4) + x_2 - (y_2 + y_4) + x_3 - (y_2 + y_4) \geq x_1 - (y_3 + y_4) + x_3 - (y_3 + y_4)$$

By, cancelling x_1 , x_3 and $2y_4$ from both sides, we will be left with:

$$2x_2 - 2y_1 - 2y_2 - 2y_4 + 2y_3 \geq 0$$
$$x_2 - y_1 - y_2 - y_4 + y_3 \geq 0$$

Now, we can see in the venn diagram that:

$$x_2 \geq y_1 + y_2 + y_4$$

as y_1, y_2 and y_4 all of them are subpart of x_2 only and is common with other sets.

So,

$$x_2 + y_3 \geq y_1 + y_2 + y_4$$

Therefore, triangle inequality holds for this distance.

So, we can tell that it is a valid distance metric.

(B.)

$$d_2(A,B) = \frac{|A - B| + |B - A|}{|A \cup B|}$$

we have to prove that

$$d_2(A,B) + d_2(B,C) \geq d_2(C,A)$$

Now by our notation we can write:

$$\frac{x_1 - (y_1 + y_4) + x_2 - (y_1 + y_4)}{x_1 + x_2 - (y_1 + y_4)} + \frac{x_2 - (y_2 + y_4) + x_3 - (y_2 + y_4)}{x_2 + x_3 - (y_2 + y_4)} \geq \frac{x_1 - (y_3 + y_4) + x_3 - (y_2 + y_4)}{x_1 + x_3 - (y_3 + y_4)}$$

Now, we can also write above equation as:

$$1 - \frac{(y_1 + y_4)}{x_1 + x_2 - (y_3 + y_4)} + 1 - \frac{(y_2 + y_4)}{x_2 + x_3 - (y_2 + y_4)} \geq 1 - \frac{(y_3 + y_4)}{x_1 + x_3 - (y_3 + y_4)}$$

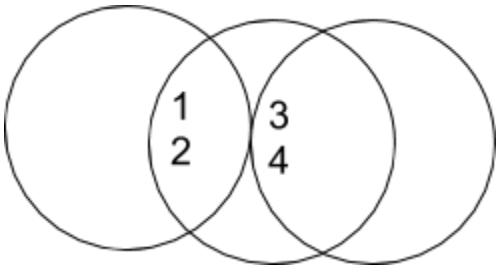
Now replacing these terms with p , q and r will result into:

$$1 - p + 1 - q \geq 1 - r$$

where p is the similarity ratio between Sets A and B and so on.
So, equation will be:

$$p+q \leq 1+r$$

Let's assume that this statement is false.
Then, $p+q>1+r$
where p is ratio of similarity between A and B. q is ratio of similarity between B and C. r is similarity ratio between A and C.



In case of $p+q=1$

Now, if $p+q$ will be greater than 1, then $A\ and\ C$ cannot be disjoint sets as Elements of B will definitely be present in $A\ or\ C$ even in case of similarity ratio sum 1. So, To achieve greater than 1, if we try to increase either p or q then $A\ and\ C$ will start intersecting and that is nothing but r. So, more we increase p or q, more r will increase. So, At best $p+q$ can be equal to $1+r$ but cannot be ever greater than it. So, our assumption was wrong. Thus, by proof by contradiction we can say that $d_2(A,B)$ is a true distance metric.

(C.)

$$d_3(A,B) = 1-\left(\frac{1}{2}\left|\frac{A\cap B}{|A|}\right|+\frac{1}{2}\left|\frac{A\cap B}{|B|}\right|\right)$$

So, we have to prove that:

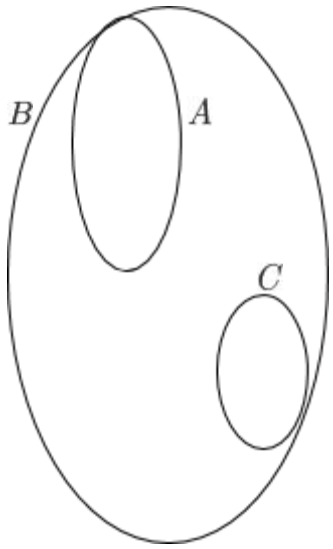
$$1-\left(\frac{1}{2}\left|\frac{A\cap B}{|A|}\right|+\frac{1}{2}\left|\frac{A\cap B}{|B|}\right|\right) + 1-\left(\frac{1}{2}\left|\frac{B\cap C}{|B|}\right|+\frac{1}{2}\left|\frac{B\cap C}{|C|}\right|\right) \geq 1-\left(\frac{1}{2}\left|\frac{A\cap C}{|A|}\right|+\frac{1}{2}\left|\frac{A\cap C}{|C|}\right|\right)$$

Now as per our notation we can write:

$$1-\frac{(y_1+y_4)}{2}\left(\frac{1}{x_1}+\frac{1}{x_2}\right) + 1-\frac{(y_2+y_4)}{2}\left(\frac{1}{x_2}+\frac{1}{x_3}\right) \geq 1-\frac{(y_3+y_4)}{2}\left(\frac{1}{x_3}+\frac{1}{x_1}\right)$$

$$1+\frac{(y_3+y_4)}{2}(\frac{1}{x_3}+\frac{1}{x_1}) \geq \frac{(y_1+y_4)}{2}(\frac{1}{x_1}+\frac{1}{x_2}) \quad + \frac{(y_2+y_4)}{2}(\frac{1}{x_2}+\frac{1}{x_3})$$

Now let's take a case,



Now putting the value of this case in the equation, we can substitute (y_1+y_4) with x_1 and (y_2+y_4) with x_2:

$$1+0 \geq \frac{1}{2}(1+\frac{x_1}{x_2}) \quad + \frac{1}{2}(1+\frac{x_2}{x_3})$$

Now,

This is clearly false as each part of the right hand side will be greater than ½. So, $d_3(A,B)$ cannot be a distance metric.

(D.)

$$d_4(A,B) \quad = \quad 1-\left(\frac{1}{2}\frac{|A|}{|A\cap B|}+\frac{1}{2}\frac{|B|}{|A\cap B|}\right)^{-1}$$

So, we have to prove that:\

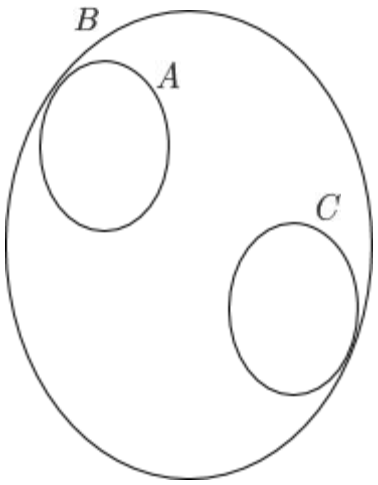
$$1-\left(\frac{1}{2}\frac{|A|}{|A\cap B|}+\frac{1}{2}\frac{|B|}{|A\cap B|}\right)^{-1}+1-\left(\frac{1}{2}\frac{|B|}{|B\cap C|}+\frac{1}{2}\frac{|C|}{|B\cap C|}\right)^{-1}\geq 1-\left(\frac{1}{2}\frac{|A|}{|A\cap C|}+\frac{1}{2}\frac{|C|}{|A\cap C|}\right)^{-1}$$

So, we can write:

$$1-\left(\frac{1}{2}\frac{|A|+|B|}{|A\cap B|}\right)^{-1}+1-\left(\frac{1}{2}\frac{|B|+|C|}{|B\cap B|}\right)^{-1}\quad\geq 1-\left(\frac{1}{2}\frac{|C|+|A|}{|C\cap A|}\right)^{-1}$$

Now, this will be:

$$1+2\frac{(y_3+y_4)}{x_1+x_3}\quad\geq 2\frac{(y_1+y_4)}{x_1+x_2}\quad+2\frac{(y_2+y_4)}{x_2+x_3}$$



Now, if we consider this case, then our equation would be:

$$1\geq 2*\left(\frac{x_1}{x_1+x_2}\right)+2*\left(\frac{x_3}{x_2+x_3}\right)$$

$$\frac{1}{2}\geq\left(\frac{x_1}{x_1+x_2}\right)+\left(\frac{x_3}{x_2+x_3}\right)$$

Now, this cannot be always be true, because in the case where $A\ and\ C$ are disjoint and A contains exactly half element of B and so is C, then right hand side would be 1.

So, there is clearly a contradiction. So, $d_4(A,B)$ cannot be a distance metric.

(E.)

$$d_5(A,B)=(|A-B|^P+|B-A|^p)^{\frac{1}{p}}$$

If we take p= 1, then $d_5(A,B)$ is nothing but $d_1(A,B)$. We have already proven that $d_1(A,B)$ is a true distance metric and this function hold the metric property. Minkowski distance tells that if any function holds the distance metric property, then we can say that

$(|A-B|^P+|B-A|^p)^{\frac{1}{p}}$ will also be distance metric for all $p \geq 1$.
Reference: https://en.wikipedia.org/wiki/Minkowski_distance.

So, By Minkowski distance we can tell that $d_5(A,B)$ will be a distance metric as well. Hence Proven that $d_5(A,B)$ is a distance metric.

(F.)

$$d_6(A,B) = \frac{(|A-B|^P+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$$

If we take p=1, then $d_6(A,B)$ is nothing but $d_2(A,B)$. Now, we have already proven that $d_2(A,B)$ is a true distance metric and this function holds the metric property. As per minkowski distance, if any function holds the distance metric property, then we can state that for any $p \geq 1$, we can tell that

$$\left(\frac{(|A-B|^P+|B-A|^p)}{|A \cup B|^p}\right)^{\frac{1}{p}}$$

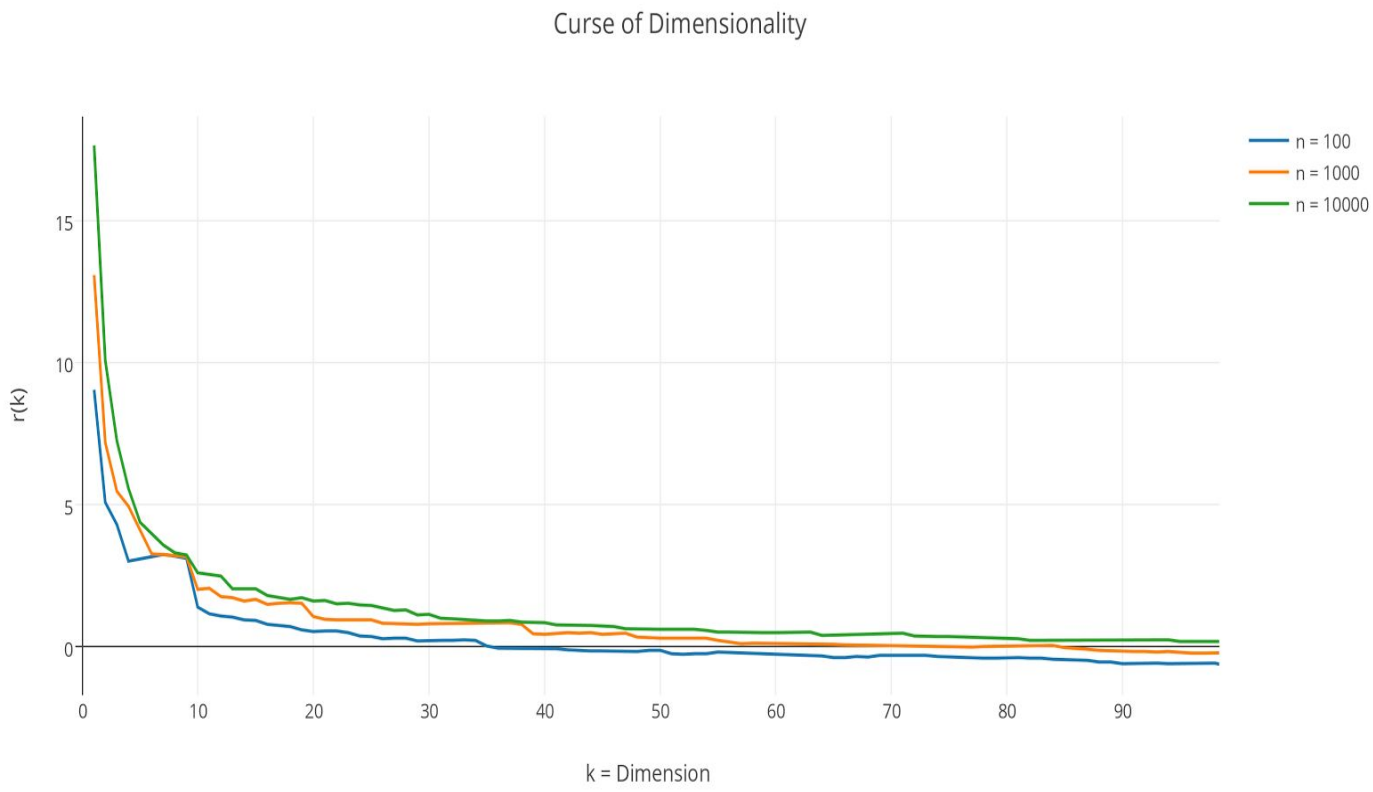
will be true.

Now taking p out from the denominator and cancelling out with $\frac{1}{p}$, we are left with:

$$\frac{(|A-B|^P+|B-A|^p)^{\frac{1}{p}}}{|A \cup B|}$$

which is nothing but $d_6(A,B)$. So, By Minkowski distance, we can tell that $d_6(A,B)$ will also be distance metric. Hence, $d_6(A,B)$ is a distance metric.

Problem: 4



Inference: From this graph, we can infer that performance decreases with the increase in number of attributes. We can also observe that with more number of data points performance improves initially but with the increase in features even it's performance deteriorate drastically. I had a notion that with more number of attributes, we would get better performance even though it would be computationally expensive but after this experiment, I realized that even performance decreases with more number of features alongside it will be really computationally expensive.

Problem 5:

Algorithm: To Implement the movie recommendation system, I applied KNN algorithms in this case. For every test case, I first extract the user information from the already stored user object and by the means of various distance formula, I find the K similar profile to the user. In first approach, I just tried various distance formula like manhattan, euclidean and L_max distance formula to identify K(as per chosen value) similar profile to the user. In second approach, I also took account of age, gender and genre to calculate the distance. After getting the distance between two user, I am taking inverse of distance to measure the similarity between two user. To this similarity measure, I also multiply with the ratio of common movie and movie watched by the user in the test case. This will give me the measure of similarity between two user.S0. by applying this formula, I rank K most similar profiles to the user in the test case. After choosing K similar user, I divide these user into four subgroups. I determine which person belong to which subgroup by applying a simple classification formula. I take the value of most similar profile and mark the similarity measure as high. Similarly, I take the Kth similar profile and mark it's similarity measure as Low. I take the difference between high and low and divide the difference by 4. Let's call this d_{25}.Now, all those users whose similarity measures lies between high and high - d_{25}, I append them in first subgroup. All those users whose similarity measures lies between high - d_{25} and high - 2*d_{25}, I append them in second subgroup. And so on, I divide all the k similar users into four subgroups. Now for each subgroup, I take the average of the ratings and multiply with the assigned weight. This score, my algorithms gives as the predicted rating of the movie by the user.

(A.) Distance_Method	K	MAD_Score
Euclidean	**50**	**0.7486**
Euclidean	**100**	**0.7386**
Euclidean	**150**	**0.7438**
Manhattan	**50**	**0.7775**
Manhattan	**100**	**0.7682**
Manhattan	**150**	**0.7734**
L_max	**50**	**0.7813**
L_max	**100**	**0.7784**
L_max	**150**	**0.7811**

(B.) I started just with the distance metric based on difference between the rating for the common movies between two user to determine the similarity between two user. Later on, I decided to include genre of the movie, the average of rating provided by different sex for the movie and the average rating given to the movie by person belonging to different age group as well to improve my movie recommendation system. Rationale behind this move was that there are numerous cases where one particular movie has been liked by any particular group. For example, It seems that action movies are more liked by the male group in comparison to the female group. On the other hand, animation movies are more liked by the person belonging to the younger age group. And it can also be the case that one user particularly like movie of few genres and do not show much interest towards movies of other genres. So. to take holistic approach to better the movie recommendation system seemed an obvious step. So, In this subproblem, I gathered the information about each movie like what is the average rating for the movie by females and by the males. Similarly, what is the average rating for the movie by the persons belonging to different age group. On the other hand, collected the information about each user that on average one gives how much rating to movie belonging to each genre. So, while calculating the distance between two user, I assigned weights to the different user depending on what is the average rating for the movie by different gender. Likewise, If the movie belongs to particular genre, then what are the average rating user gives to the movie of these genres. In case of multiple genre, I take weighted average. On the same line, I also assigned some weight to the difference between age of the users to calculate distance. On this basis, I select K most similar profiles in second case . Afterwards it is more like the subproblem A only.

(C.) For this case, there were some problems with the test cases. As users in the test cases were not present in the training data. So, I wrote a script to randomly generate the test case. On this generated test case, I got the following result:

Distance_Method	K	MAD_Score
Euclidean	**50**	**0.7134**
Manhattan	**50**	**0. 7386**
L_Max	**50**	**0.7483**

I did not get the time to try different K for this case. As, it takes too much time to complete one run.

Observation: For small training data, KNN gives a decent performance, but when size of training data increases, then KNN becomes too computationally expensive. It takes too much of computation cycle to determine K similar users among large number of users. So, for large

dataset, KNN is not a good algorithm in the present form. We might do some changes to quickly find K similar users. with the large dataset, it is more likely to discover some high level pattern which cannot be feasible in the case of small dataset.

(D.) Improvement Scope: I genuinely felt that the movie recommendation system could be made better by trying and testing various methods. One way I was thinking that we might improve the system by taking into account the imdb rating of the movie as well in consideration for the distance measurement. Similarly, I did not take any advantage of the zip code of the user which were provided to us. It might be the case that persons from nearby places prefer particular kind of movie. Because of lack of time, I was not able to check large number of parameters for the age, gender and genre. I feel that given appropriate weightage to each of these parameters, we might get more accurate predicting system. So, these might be the some of the steps that one can take to improve this movie recommending system.