**Similarity Functions**

Similarity functions or similarity measure is function which is used to calculate the similarity between two objects. They usually take two objects as input, and give a large positive value for the objects with high degree of similarity and either negative value or zero for the objects with low degree of similarity. It can also be defined as the distance between the dimensions or features representing the objects. The similarity functions find applications in recommender systems, clustering, sequence alignment etc.

Similarity is generally measured in the range from 0 to 1. Similarity equals 1 if two objects are identical and 0 if two objects are completely different. There are various similarity functions which are being used these days, few of them which we considered for our experiment are:

1. Euclidean Distance
2. Manhattan Similarity
3. Minkowski Similarity
4. Cosine Similarity
5. Jaccard Similarity
6. Dice Similarity

**Euclidean Distance**

Euclidean Distance is described as the distance between two objects in the Euclidean space. Euclidean distance works well on when data is dense and continuous. It can also be referred as the length of path connecting two objects. This distance is calculated using Pythagoras theorem. The function for Euclidean distance is given in equation 1.

 (1)

**Manhattan Distance**

Manhattan Distance is a metric in which the distance between two objects is calculated as the sum of absolute differences between their Cartesian coordinates. In simple words it is the absolute sum of differences between the coordinates of two objects. The function for Manhattan distance is given in equation 2.

 (2)

**Minkowski Distance**

Minkowski Distance can be considered as the generalized metric form of Manhattan Distance and Euclidean Distance. The function for Minkowski distance is given in equation 3.

(3)

The most important parameter in the above equation is c. If c = 1, then the equation behaves as Manhattan distance and if c = 2, then the equation behaves as Euclidean distance. Similarly, when c equals infinity the resulting equation is called Chebyshev distance.

**Cosine Similarity**

Cosine similarity is the measure of similarity between two vectors which finds the normalized dot product of two vectors. By trying the calculate the cosine similarity we are effectively trying to figure out the cosine of angle between two objects. If two vectors have the same orientation then their cosine similarity is 1, if two vectors are at right angle then their cosine similarity is 0. The cosine similarity is generally used in positive vector space, where the result is bounded by [0,1]. The function Cosine Similarity is given in equation 4.

 (4)

**Jaccard Similarity**

Jaccard Similarity or Jaccard index is a similarity measure which is used to calculate the similarity between two finite sample sets. It can be defined as the cardinality of intersection of two set divided by cardinality of union of two sets. The function Jaccard Similarity is given in equation 5.

 (5)

**Dice Similarity**

Dice similarity is also known with several other names like Sorensen index and Dice coefficient. It is used to calculate the similarity between two samples. Dice similarity is very similar to Jaccard Similarity. It also doesn’t satisfy the triangle inequality property just like Jaccard Index. Dice similarity works well with heterogeneous data and gives less weight to outliers hence reducing the effect of outliers on the final output. The function Dice Similarity is given in equation 6.

(6)