Read me file: - we have created this project in a group. Following are the names and ids of the students in the group: -

* Udbhavsinh Solanki (0873798)
* Jay Patel (0872908)
* Bhavya Faldu ( 0872906)

Similarity Analysis

Here, I try to calculate Jaccard similarity, Cosine Similarity, Euclidean distance between the set of docs given.

Initially, we calculate the tf-idf for the documents given.

**What is tf-idf?**

As by the word, it stands for term frequency - inverse document frequency. It is statistical value, used for calculating the importance of a word in the document in the corpus( a collection of documents is called corpus).

tf-idf is composed of two words. First one is term frequency and the second is inverse document frequency. we calculate both of them separately. Term frequency is nothing but the number of times a term appears in the document divided by the number words in the document. whereas inverse document frequency is logarithm to the total number of documents in the corpus divided by the number of documents the term actually contains.

Another important term that we use is ‘stop words’. what is stop words? stop words are nothing but commonly used words like the,a,is etc., These words are given little value so we exclude them from documents after we tokenize the words.

**Coming to our similarities**:

**What is Jaccard similarity**?

It is the measure of the total number of common words in both the documents divided by total number of words in both the documents. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It’s a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations. Although it’s easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations.

What is the Euclidean distance?

* Here it is the Euclidean distance between two documents represented as vectors. The distance between two objects that are not points is usually defined to be the smallest distance among pairs of points from the two objects. Formulas are known for computing distances between different types of objects, such as the distance from a point to a line.

**What is Cosine Similarity**?

Here we represent each document as a set of vector. let n be the total number of words in all the documents or corpus. Then each document is represented as a vector of dimensionality n. Then we find the cosine value between two documents using these vectors.