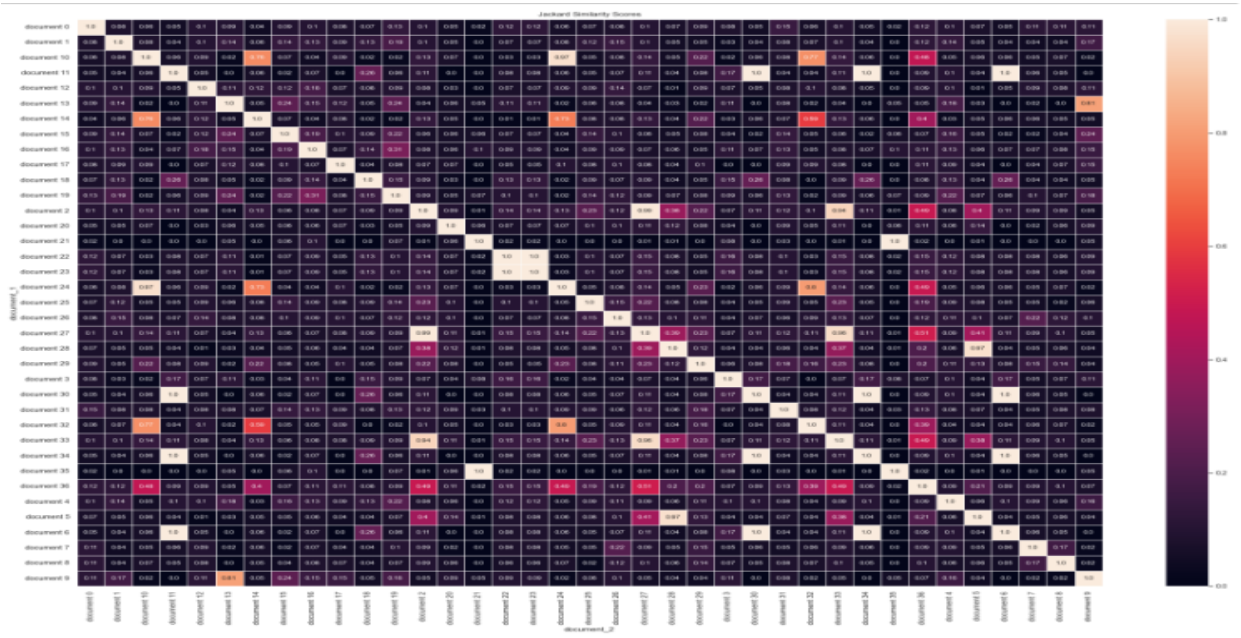
ML Lab 02 Report

Name: Raman Kumar
Reg. No: BL.EN.P2DSC22009
Course Name: MTECH Data Science

Q.1 What are your observations from the similarity study conducted in this experiment?
Please design a plagiarism detection system based on your learnings in this exercise. Mention the advantages and disadvantages of your designed system with respect to performance, efficiency and accuracy. Please highlight failure scenarios of your designed system.

Measuring similarity is actually finding how much similar or closer the data points are to each other. We measure these similarity experiments in numerical values, the higher the similarity value will be the more closer or related the elements will be to each other. From the experiment done on the Reports data, I observed that the the reports submitted by the students is somewhat in the words of their own.
I also visualized the data with heatmap also that gives me the exact values in those (37, 37) matrix or array. There were total 2906 words or tokens in the dataset. We reduced them to 466 tokens by using the stopwords whose count was 179.



By observing the heatmap, there are much black region/cells. The black region indicates the dissimilar words in the documents. The red region indicating the document with higher plagiarism or similarity.

We used cosine and Jackard similarity here to get the similarity in the documents.
In *cosine* similarity, we are counting the number of 1's i.e positive matches and taking the square root of them. In cosine if the value of the vectors will be more than the similarity will be more, as a result the plagiarism will be higher.
In *Jackard* similarity, we are measuring the distance in between the words i.e arithmatically. If the document will be so near to '0' then the dissimilarity in between them will be high. In Jackard if the dataset will be so high most of the times it will be near to '0'.

- D(A, B) = 1 - J(A, B)

In plagiarism detection system I trained a language model based on the reports document. I used the nltk library to implement it. I read the document file and split the 'Reports' column from it. In 2nd step I removed the punctuations and removed the stop-words. In next step I vectorized the sentences/words. Then I experimented with some distance/similarity measure methods. Compared each word with the source file documents. Then the words passed through a threshold value. On getting the value more than the threshold value the result displayed as 'Similar Words'.

Advantages: These tools are mainly used in finding the piracy or stealing of data from various famous papers and research works. Also in education purposes they are using these to check the assignments and exams.
Disadvantages: There are so many different words that we can't remove from the bag. Also some words are most common words those are not actually present in stopwords so that will lead to plagiarism. And the system might also reject the words which have length greater than 8-10.

Q2. Mention a few use cases where similarity measure is useful for classification / categorization and may be used instead of distance measure.

- Plagiarism check/detection from the various documents or text based data is the real use case. It classifies the positive or matched words and rejected or dissimilar words.
- We are use distance measure in one of famous classification algorithm k-NN classifier where we use the euclidean distance to classify data points and helpful in finding the space for new data points in later trainings.
- Similarity measure has a real time use in classifying the clustered data. In clustering we don't have any prior labeled data to classify so distance or similarity measures are the only options remaining.
- One real use case of similarity measure especially cosine is retrieval of information from the internet in form of text, email etc. We can make the vector

norm of the data and compare them with cosine to filtering out the unstructured elements.