**REPORT**

**Brief introduction about dataset:**

The data contains a pair of paragraphs. These text paragraphs are randomly sampled from a raw dataset. Each pair of sentences may or may not be semantically similar.

**Problem Statement:**

The candidate is to predict a value between 0-1 indicating the similarity between the pair of text paras.

1 means highly similar

0 means highly dissimilar

**Solution:**

For finding the solution we need to import basic python libraries required.

i.e. numpy and pandas.

After importing the libraries we read the dataset using pandas..

Reading the dataset we came to know that there were two columns and three thousand rows present in the dataset.

Then we found any null values present in the dataset using python and we came to know that there was no null values present in the dataset.

After that we did preprocessing of dataset. Steps involve in preprocessing are as follows:

* case conversion :- To convert para into lowecase or uppercase.
* word tokenization:- Break the sentences into tokens. (words)
* non-aplha removal :- Remove non alphabetic character.
* Stop-word removal :- Stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. In NLP and text mining applications, stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead.
* Lemmatization :- It is used to normalize words into its base form or root form.
* Joining of list of msgs.

After preprocessing the dataset did the vectorization and finally found similarity between the two texts using **cosine similarity** function of python.

Then made web application for user by using python flask framework.

After that deployed the algorithm on cloud server.