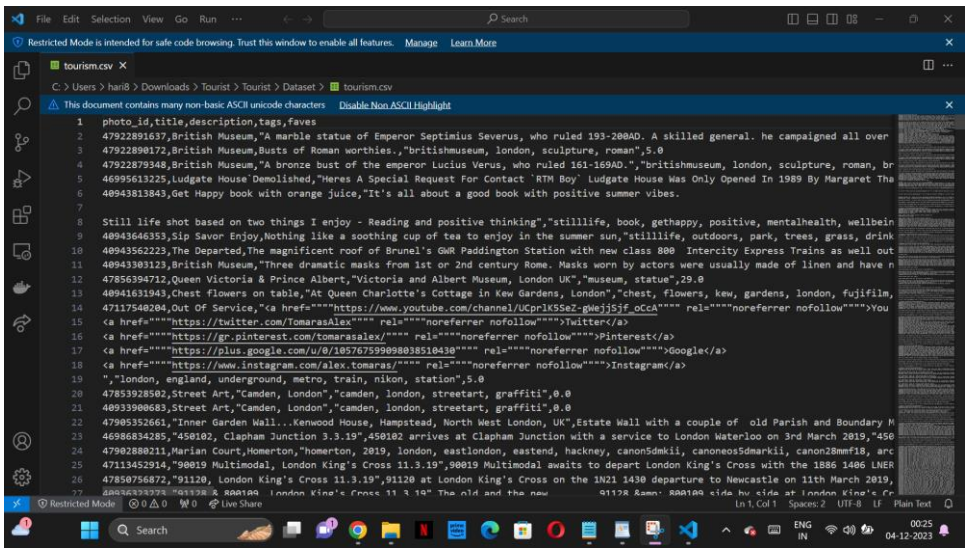
Analysing Tourist Behaviour using Big Data Technology

In this project we are implementing Popular Tourist Place Recommendation using Bigdata framework called SPARK. Spark can process data in distributed manner so it can handle any size of data. In propose work we are using Geo Tagged images dataset to extract places information and then this text data will be cluster to group similar behaviour tourist into same cluster. Whenever new user input any query then clustering algorithm will predict similar cluster as per user query and then suggest top 5 popular visited similar interest places as recommendation.

To cluster user behaviour we have used KMEANS algorithm and then the user features and cluster label will get trained with Random Forest Classifier to explain model and for explanation we have used SHAP framework instead of LIME. This model will explain which features contribute most to predict particular label.

For this project we have used below tourism data which is Geo Tagged using FLICKER

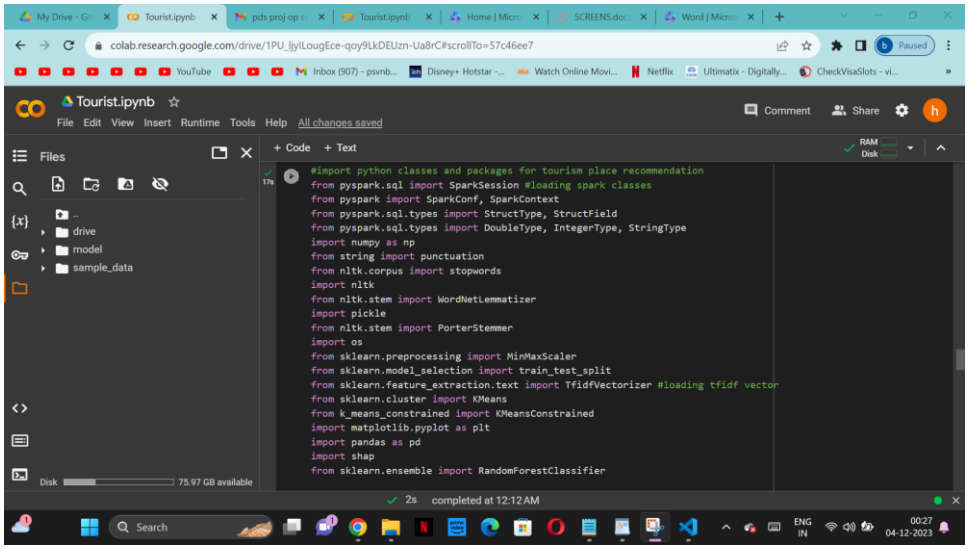


In above dataset first row represents column names and remaining rows represents dataset values and in columns we have names as “Photo ID, Description, tags, favourites as Number of time visited etc.”. So by using above dataset we will cluster and recommend places for new user.

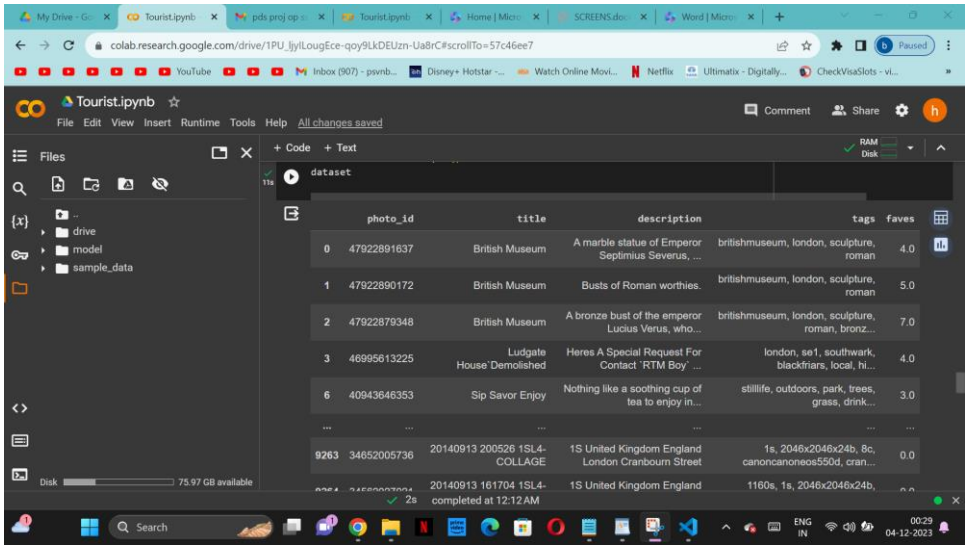
SCREEN SHOTS

We have coded this project using GOOGLE colab and below are the code and output screens with comments

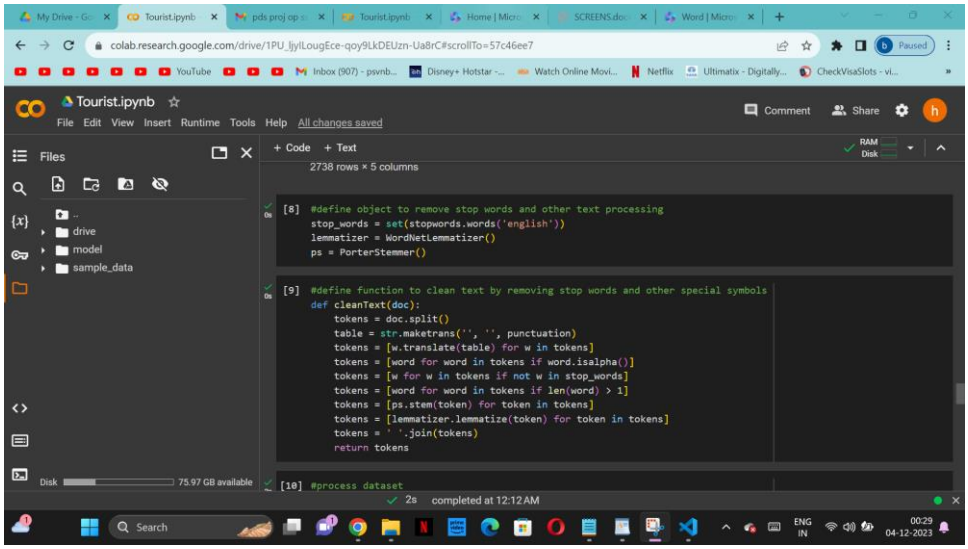
In below screen importing spark, NLP (natural language processing API to remove stop words, special symbols from geo tag text dataset) and then importing KMEANS and other classes



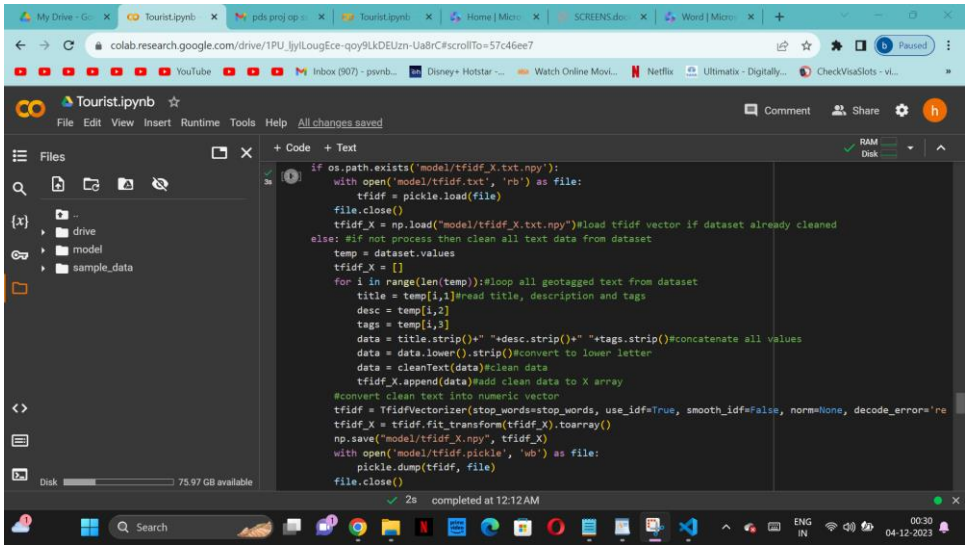
In below screen using Spark we are loading Tourism dataset and after loading will get below output



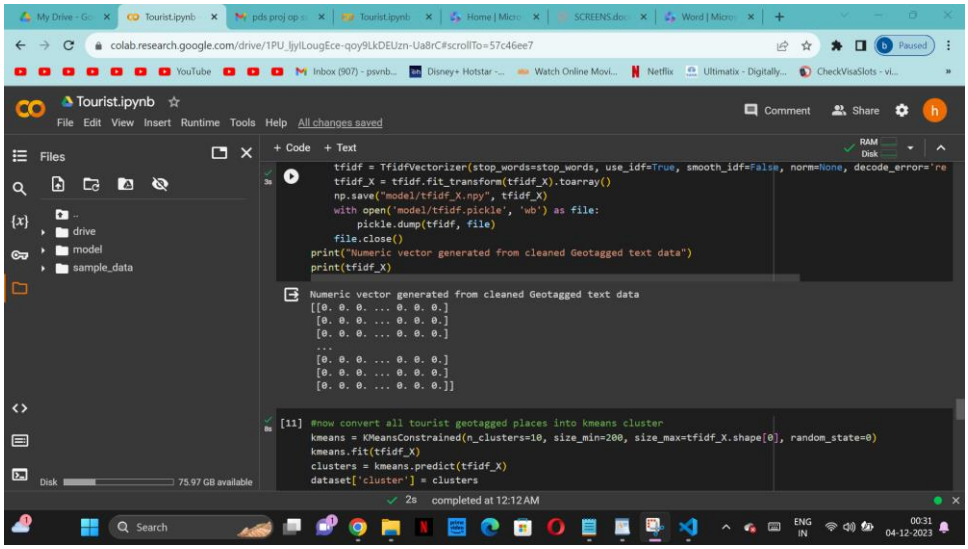
In below screen we can see loaded dataset values



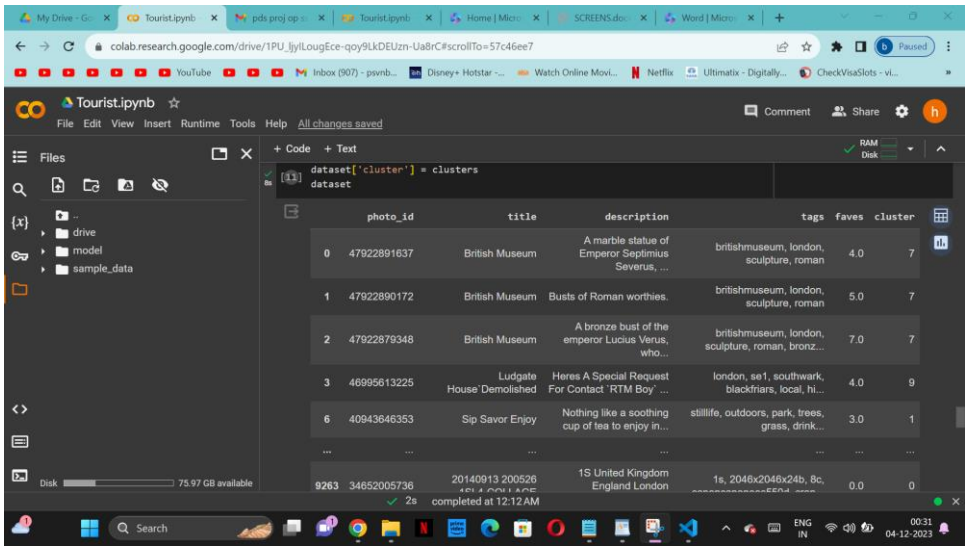
In below screen defining function to clean geo tagged text data



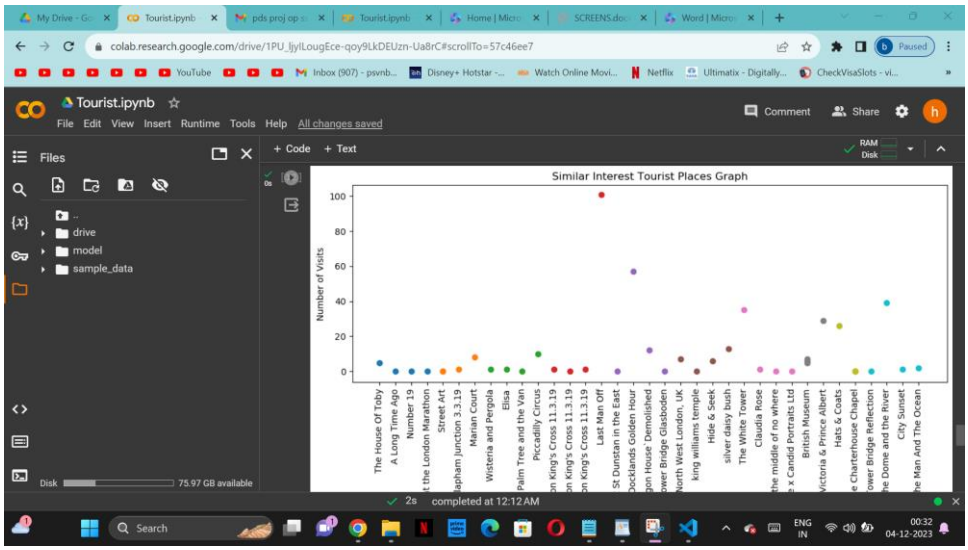
In below screen reading all Geo Tagged text data and then converting all clean text data into numeric TFIDF vector and this vector contains average frequency of each words and if word does not contains then vector will have 0 and by using this vector KMEANS will perform clustering



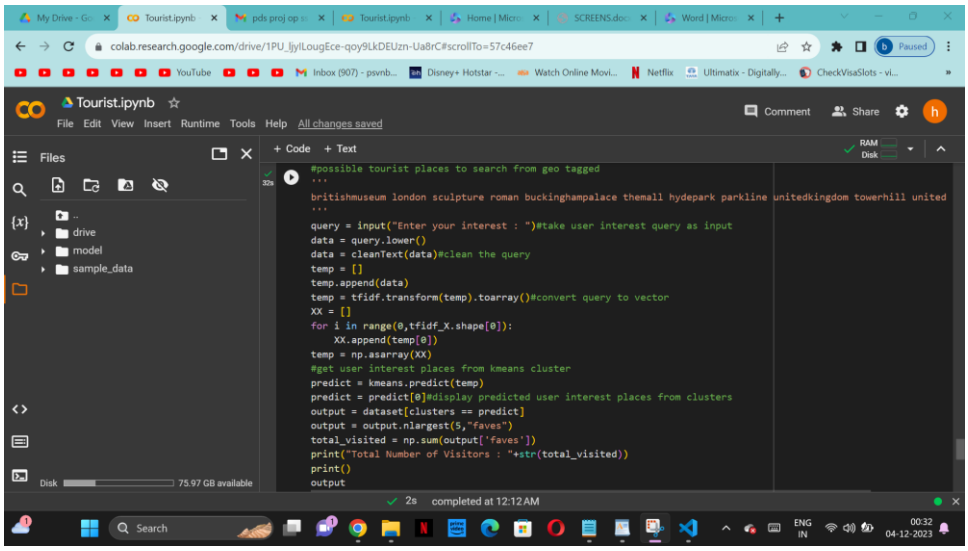
In below screen displaying TFIDF vector for few rows from dataset



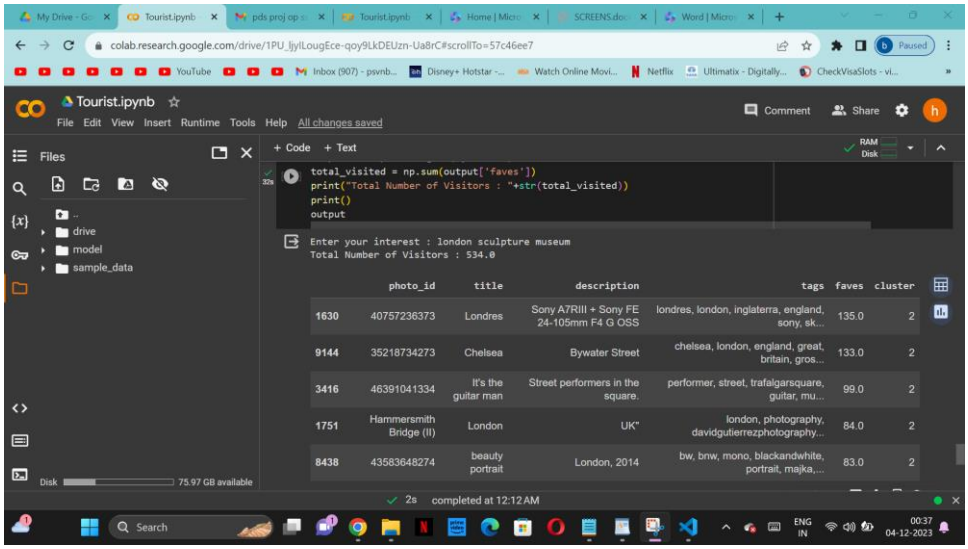
In below screen we performed clustering on all text data and after clustering in output last column we can see which row or user place goes to which cluster. In above table in last column we can see cluster label for each records. In above clustering we have created 10 clusters so all rows will be distributed between 1 to 10 clusters.



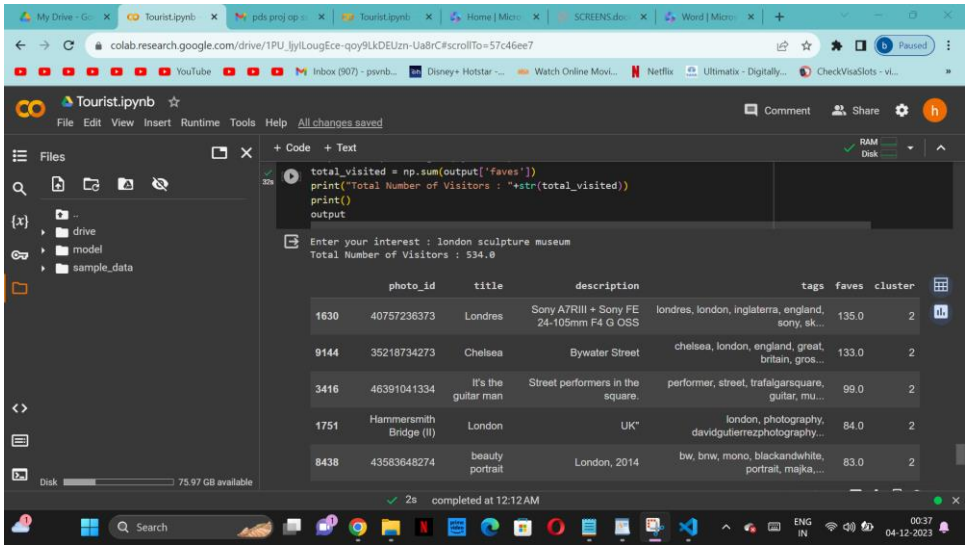
In above graph x-axis represents place names and y-axis represents Number of visited and small dots represents number of time place is visited



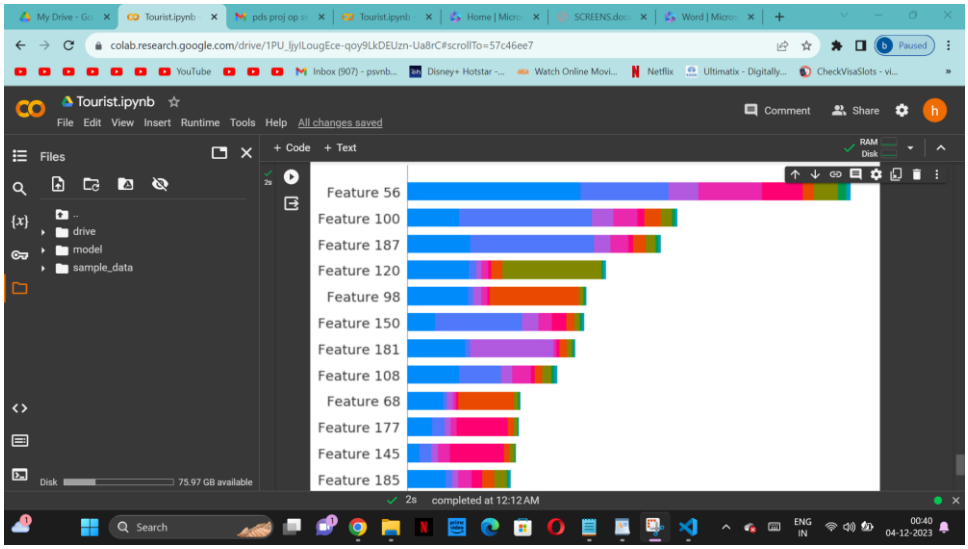
In above screen we define function which will read user query and then recommend similar places from cluster based on user interest query

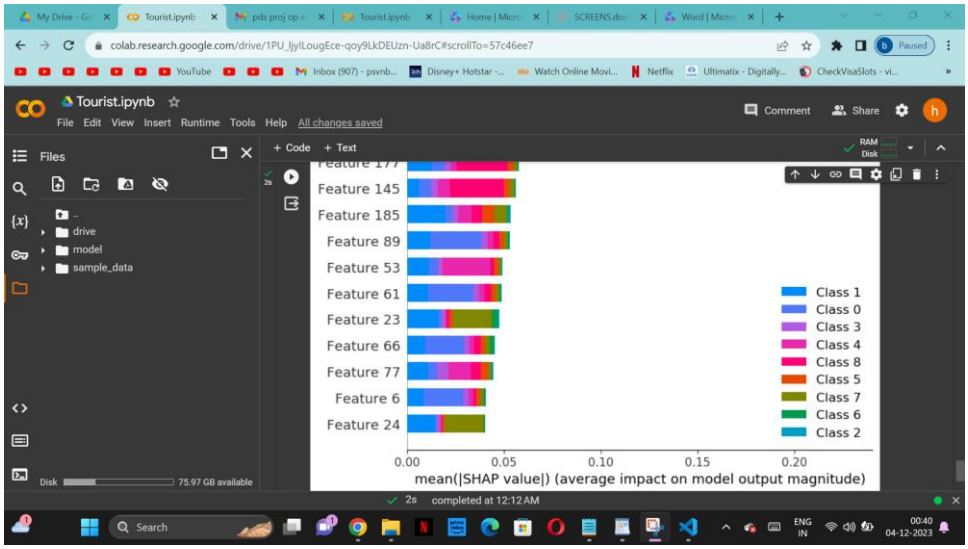


In above screen in Text Box we gave query as ‘london sculpture museum’ and press enter key to get top 5 recommended places like below screen



In above table we got top 5 recommended places from cluster 2 and in blue colour text we can see number of users visited those places. Similarly you can enter query and get popular tourist places from cluster. Above output can be consider as future tourism places which will be in demand





In above screen we have added SHAP modelling tool to explain about the model who is using which features most for prediction.