**README**

Sections :

* Code Base Folder Structure
* Database Google Drive Links
* Result Graphs Google Drive Links

**Code Base Folder Structure**

|----Code\_Submissions

|---Codes

|

|---- CountBuildings.py : Contains code to get the number of buildings in each satellite image

|

|---- UnetSegmentation.py : [**tensorflow**] : UNet model trained, and used to predict image segmentation

|

|---- PredictionsAreaCoverage.py : [**recommendation**] : Used to predict the green and urban land area coverage

|

|---- SimilaritySearch

| |---- finalProjSimilaritySearch.py[**Spark, similarity search**]: min-max scalar, finds all pair cosine similarity, filters top 100 highly correlated users(countries with year), imputes the missing values, converts rdd to csv again for hypothesis testing.

| |---- data: reads the dataset from here, saves all the heavily one-time computation req’d rdds to pickle to save time and effort, saves the newly imputed csv.

|

|--- HypotheisisTesting.py [**Spark, Hypothesis Testing]:**

| Used to generate null hypothesis on the target attributes.

| Test the hypothesis using multiple linear regression

| Verify if we can reject the null hypothesis or not

| Obtain the attributes on which our target attribute is depended

| Returns JSON file to generate recommendations for the target attributes.

|

|---- ARIMAX\_time\_series\_forecasting.py:

| Used to forecast values for target attributes for hypothesis testing

|

|---- Interactive Visual User-Interface:**[D3.js, Python, Html, CSS, Flask]**

| |----index.html: html element to create divs and input elements like dropdown. Toggle, radio button. Contains the view if the UI

| |---- main.py: does preprocessing, creates MDS plot with given dissimilarity matrix(correlation), finds number of clusters(K-means, elbow plot method) and provide the client-server support to the application.

| |---- static:[**.Js files and css**]

| |---- creates all UI Plots(Barcharts, GeoMap, Stacked TIme chart, MDS, PCP) with all the brushing and linking events from any dashboard to others.

| |---- style.css: contains all the css styled classes

|

**Datasets Used** :

The datasets and the files obtained after preprocessing can be accessed on this google drive location

<https://drive.google.com/drive/folders/1KHL9X2qNMjKFelu9TCSoLLqKnYetEU5y?usp=sharing>

( can be accessed through Stony Brook Computer Science Department email)

**Results Obtained** :

The results obtained in the form of graphs and trained models, have been placed in this google drive location

<https://drive.google.com/drive/folders/13hvO86XGdL-R_qbigLf5iVzcCdsRW9jT?usp=sharing>

( can be accessed through Stony Brook Computer Science Department email)

For Complete set of code on Similarity Search, Interactive Visual Dashboard, and Hypothesis testing, please look at the github repo: <https://github.com/dead-pool-kit/CSE-545-final-project> (Pulkit Varshney)