# README - 590R Assignment 4

Submitted By :- Ankita Mehta

November 27, 2017

This document will take you through the implementation of document clustering performed on shakespeare dataset . It has been written in Python language. This part has been build over the previous submission of VSM Retrieval Model. To run the project, change the current working directory to Assignment4/src. Paths mentioned in the commands written below are according to the above current working directory.

**NOTE: python RetrievalAPI.py -h commands will give you the information about the command line parameters of the wrapper.**

## 1 Code Dependencies

Python 2.7.12 :: Anaconda has been used for this project.
Pip version 9.0.1
Installing argparse: pip install argparse

## 2 Build and Run the Code

**Code structure** :
The code structure has been divided into three main folders: Assignment4/data, Assignment4/src, Assignment4/results.

*Assignment4/data:

1. shakespeare.json: It is the input data file for creating compressed/uncompressed index
2. unc_manifest: : It is the manifest file having the names of files and indexes created for uncompressed file.
3. comp_manifest: It is the manifest file having the names of files and indexes created for compressed file.

*Assignment4/src:
This folder has some codes from the previous assignment :

1. main.py: It is the entry point for finding the indexes, generating query files with unique random words and dice coefficient.
2. indices_creation.py: It is the class having functions for finding the indexes, generating query files with unique random words and dice coefficient.
3. Encoding_decoding.py: It is the class having functions for performing delta and Vbyte encoding-decoding.
4. API_extract_statistics.py : It is the class to extract vocabulary, Collection Term frequency and document frequency for the query word.

Codes which have been amended/written specifically for this assignment are :

1. clustering_wrapper.py : It is the entry point for implementation of agglomerative clustering.
2. agglomerative_clustering.py: This code implements agglomerative clustering on the documents present in the shakespeare dataset.

3. linking_and_Cosine_similarity.py : This code implements various linking choices - min, max, average, mean and also finds cosine similarity between two documents using VSM retrieval model.

**Note:** In all clustering experiments, cluster numbers have a range from 1 to 748 whereas document numbers are starting from 0 to 747.

*Assignment3/results: It has some previous results (6 for each) for compressed and uncompressed index to be used for this assignment.

1. xxx_docNo_playId: It is the mapping from doc_No to playId
2. xxx_docNo_sceneId: It is the mapping from doc_No to scene_Id
3. xxx_lookup_table: It is the lookup table having mappings from term to doc_No , count , Collection_term_frequency , document_frequency
4. xxx_sceneId_docNo: It is the mapping from scene_Id to doc_No
5. xxx_Inverted_list: It is the binary file having stored inverted lists.
6. xxx_docNo_length: It is the mapping from doc_No to its length.

Note: Here xxx is either 'unc' or 'comp' for uncompressed and compressed index respectively.

Other results relevant to this assignment are:

1. cluster-< *thresh* >.out : Results have been saved for threshold range 0.05 to 0.95 with step size 0.05 using mean linkage.

The other files present in the Assignment4 folder are : README, report.pdf

## 2.1 clustering_wrapper.py

One wrapper have been written for this project i.e. - **clustering_wrapper.py** having location: Assignment4/src

```
python clustering_wrapper.py -m ../data/unc_manifest
```

**-m** specifies the manifest file path - the file having all paths written

**Note**: There is just 1 variant required to run this code and it performs clustering for threshold range 0.05 to 0.95 with step size 0.05 and using four types of linking choices : min , max , average and mean. However, results are just taken with mean linkage choice for all thresholds.