**Readme document for Multi Viewpoint Based Clustering implementation**

The main input for this program comes from the index.html file that lists all the other files that are to be considered for clustering.

Each html file needs to be generated in a particular format as can be seen in any of the html files used in the current demo.

All generated html files needs to be kept in a single folder and then an index.html needs to be generated which contains links to all these files. Index.html can also contain data that is part of the clustering process.

Please find the link to the JDK bin folder on your system and replace the path variable value in mvcs.bat.

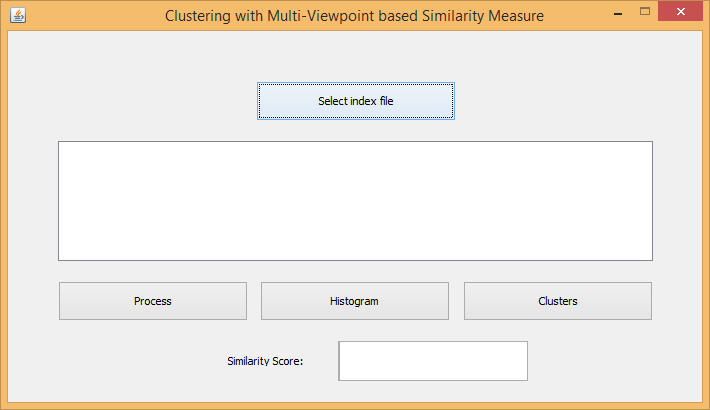
If you do not have Java JDK installed you can download from here:

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

Please find the path to the folder where you have kept the code and replace the cd value in mvcs.bat with that value.

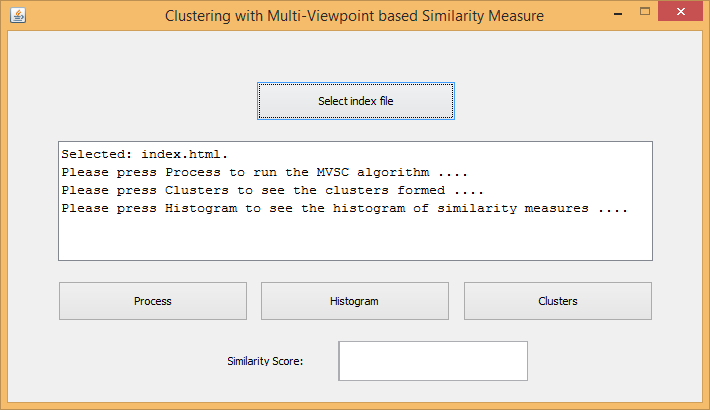
Please launch the batch file by right clicking on it and executing the administrator mode. Accept all security prompts that are thrown. Double clicking will also work.

The landing screen looks like the following:



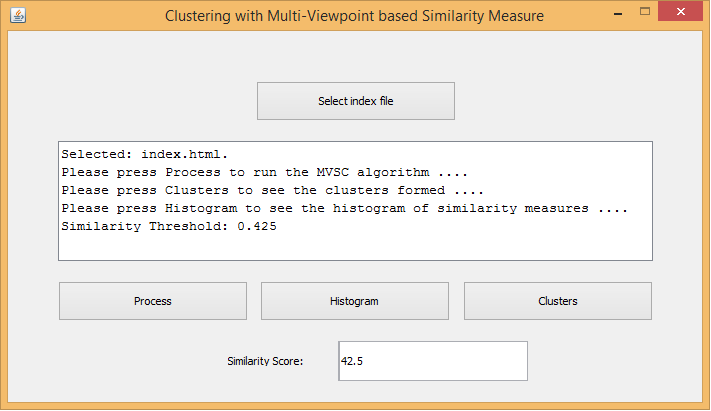
Hit Select index file to navigate to the index file which links to the rest of your dataset and select it.

After this selection the window will look like this:

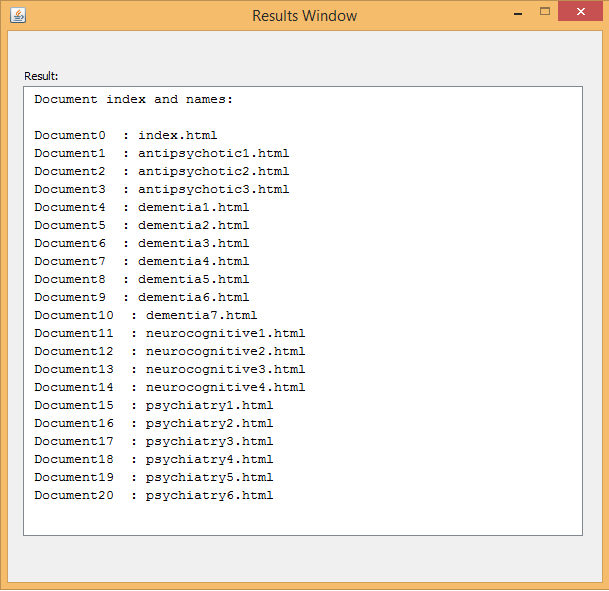


The similarity score threshold for clustering is set at 30 (0.3) which will be the default value if nothing is specified in the Similarity Score input box.

If other similarity score is needed then you should specify it in this box and press enter. A confirmation message will show up in the display box as follows:

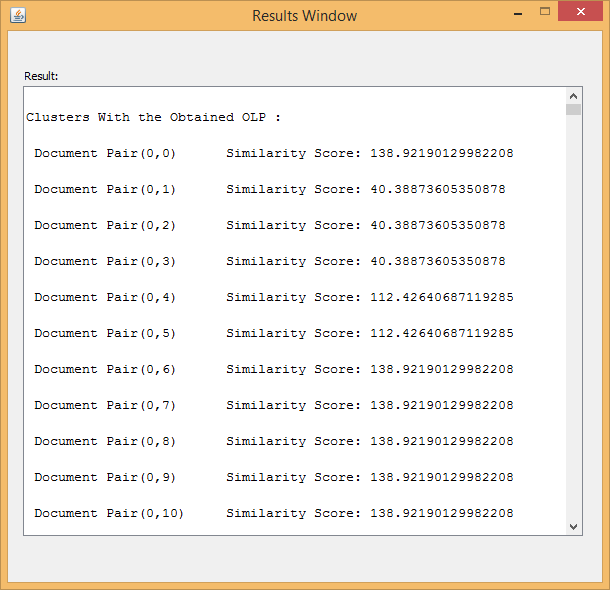


At this point please hit the Process button to process your files. At the end of processing the results window will show up as follows:



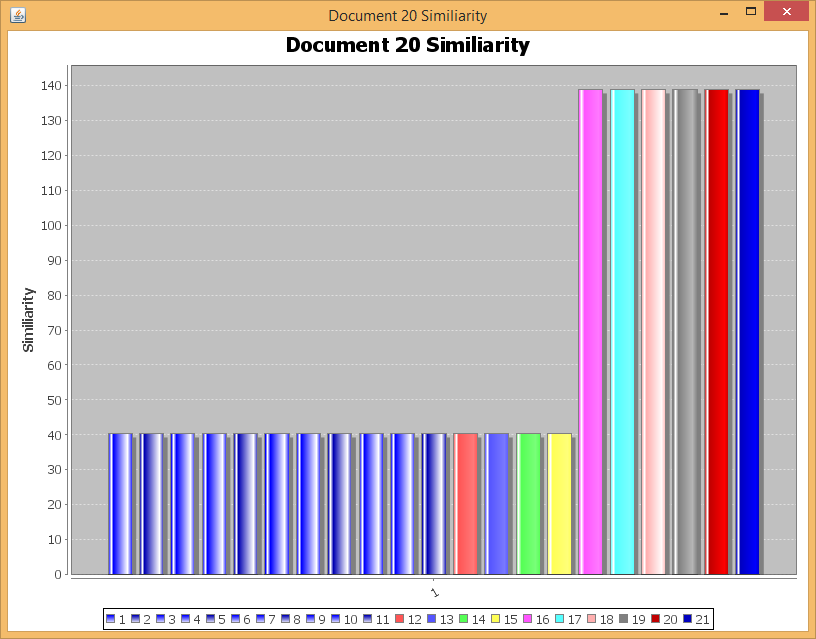
This gives us the document indexes which will be referred to in the clustering process and showing of the histogram. This is saved in your code folder as a process\_log with the run number and the timestamp.

At this point you can hit clusters to view the document pairwise similarity scores, then the clusters formed for the specified similarity score threshold and the number of clusters formed. This data is also saved in your code folder as a cluster\_log with the run number and the timestamp.



Now you can also use the similarity threshold value and hit clusters again for re-clustering based on the new parameters. Every such run will be saved in separate cluster\_logs.

Histogram can be also seen now on hitting the histogram button. A histogram is thrown for every document showing it in comparison to all the other documents by similarity score. A sample is provided below:



If a new dataset needs to be processed now then we need to select a new index file and repeat the above. A new process\_log will be saved every time the Process button is pressed.

The methodology of implementation of the idea is as follows:

1 . Raw Data is taken and formed into structured html documents with assigned key words that the user will decide.

2 . Parsing of the html files and constructions of document index graphs from the key words.

3 . Phrase and term matching by TF and IDF measures.

4 . Clustering based on the similarity scores derived.