README

The project is a Search Engine written in Python that uses the Vector Space Model and TF-IDF scoring for information retrieval.

The 20newsgroups data set has been used, and it has approximately 15,000 documents. The data set used can be downloaded in a zip file from :

<https://drive.google.com/file/d/0B5IBLP79305iMUs4OFpURkYyS00/view?usp=sharing>

For using this or any other data sets, download and keep the data set in the same directory in an unzipped folder that contains text files whose names are **just sequential numbers**. Ex – 1, 2, 3, 4, etc.

Specifically in this case, unzip the provided file “Files.zip” in a folder called “Files” in the same directory where the code and this README are present.

Open this Readme file’s working directory in command line and type “*python 106\_ir.py*” to run the code. It has a few dependencies such as NLTK which can be easily pip installed.

You will be prompted to enter your query after a while. At this point type your query into the terminal and wait for the results to be displayed on the screen.