The project has three parts:

1. Compute the page rank for UTD CS department web pages.
2. Compute TF.IDF for web pages from UTD CS department
3. Build a query engine which returns the top web pages for the user specified queries.

Part 1:

Python is used for web crawling and also for web text retrieval.

1. Run python run\_script.py

This will generate the following files as output:

1. prank.txt – flattened adjacency list of all web pages where web pages are mapped to unique integer
2. urllist.txt – List of all unique webpages associated with unique integer
3. It also creates a directory called “webText” which will have text data of all the web pages. One file is created per web page.
4. Copy prank.txt and **urllist.txt** to HDFS which will be input for page rank algorithm
5. Run the following command from UTD cluster where page\_rank.scala is present. It takes command line argument which is the absolute path of prank.txt and urllist.txt and the output directory.

spark-shell -i page\_rank.scala --conf spark.driver.extraJavaOptions="-Dhdfs:///user/ass150430/prank.txt, hdfs:///user/ass150430/urllist.txt, hdfs:///user/ass150430/pageRank " --master yarn

1. Store the output in a single file as **pageRank.txt** which will be used by Query Engine.
2. Copy webText directory to HDFS which was created in step 1).
3. Run the following Hadoop jar command to generate inverted Index. It uses webText as an input directory and it also takes numberOfDoc as a command line argument. It’s value should be the number of lines/entries in urllist.txt

hadoop jar InvIndexIDF-0.0.1-SNAPSHOT.jar invIndexIDF.invIndexIDF.InvIndexIDF -D numberOfDoc="445" -D stop\_file="/user/ass150430/stop-word-list.txt" /user/ass150430/webText /user/ass150430/iIndex

hadoop jar InvIndexIDF-0.0.1-SNAPSHOT.jar invIndexIDF.invIndexIDF.InvIndexIDFPair -D numberOfDoc="445" -D stop\_file="/user/ass150430/stop-word-list.txt" /user/ass150430/webText /user/ass150430/iIndexPair

1. Store both the outputs in a text file as **iIndex.txt and iIndexPair.txt** which will be used by Query Engine
2. Copy the Query.jar into any directory along with iIndex.txt, iIndexPair.txt, urllist.txt and pageRank.txt and run the following command for the interactive query engine:

java -jar Query.jar invIndexIDF.invIndexIDF.QueryEngine

