Web Search Engines – Project – Abhishek Bhunia (ab5966)
A dynamic, in-memory Document Clustering system

A) Instructions to run the program on the server:

(**hosted and tested on linserv1.cims.nyu.edu**; source code also uploaded on classes)

1. Download and extract the source and libraries. Everything should be extracted in one folder, maintaining the folder structure of the archive.
2. Compile ***from*** the extracted folder:

   javac -classpath ".:./lib/*:./lib/carrot/required/*:./lib/carrot/optional/*" *.java

```
ab5966@linserv1[cgi-bin]$ pwd
/home/ab5966/public_html/cgi-bin
ab5966@linserv1[cgi-bin]$ javac -classpath ".:./lib/*:./lib/carrot/required
/*:./lib/carrot/optional/*" *.java

ab5966@linserv1[cgi-bin]$
ab5966@linserv1[cgi-bin]$
```

3. Run ***from*** the extracted folder:

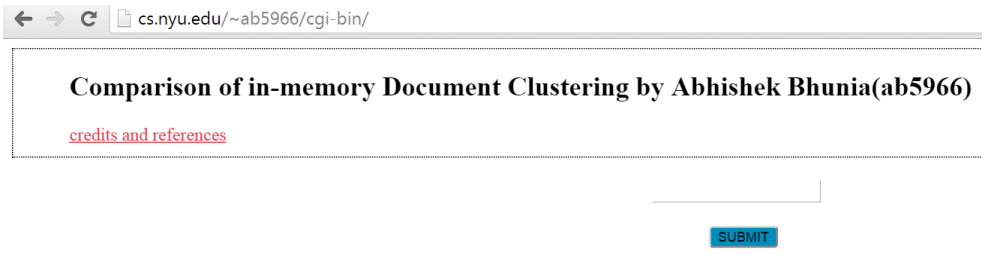
   java -classpath ".:./lib/*:./lib/carrot/required/*:./lib/carrot/optional/*" MainApp -index /home/ab5966/public_html/cgi-bin/index -docs /home/ab5966/public_html/cgi-bin/crawldocs -query "blue whale"

   The index and crawled document folders ***have to be present***

4. Successfully tested queries: learn dchinese^, blue whale, king cobra, brad pitt, why mercury is so hot?, spacex

B) Instructions for testing through the Web App: *(Please read special note in Page 3)*

1. Go to: http://cs.nyu.edu/~ab5966/cgi-bin/

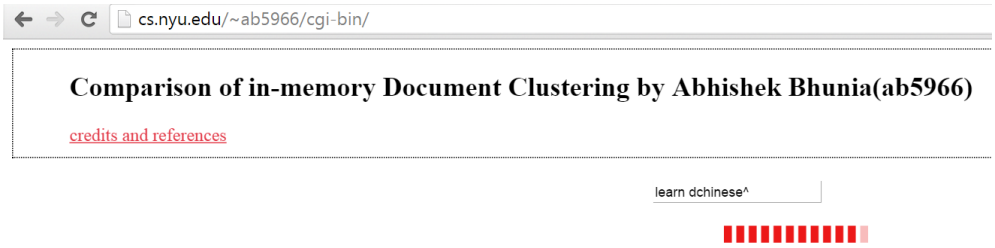


2. Submit your query and wait for about 30 seconds. It takes so long because everything from making Bing API calls, retrieving web pages, making a lucene index, running three clustering algorithms on the retrieved pages takes place at run-time and in memory.

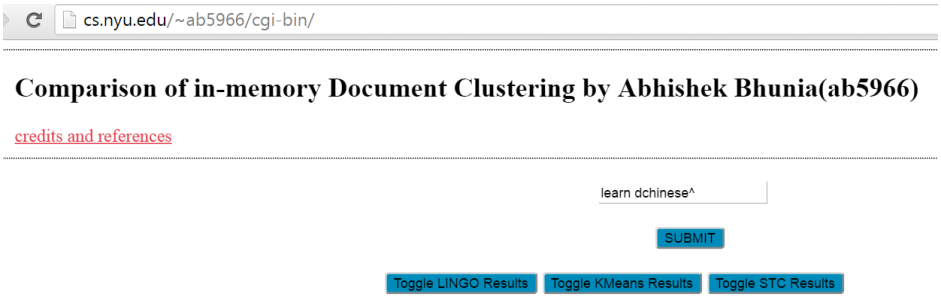
   I explained this response delay to Professor and the actions I took to minimize it and he is okay with it –

   I) Got rid of my own crawler and replaced with Bing API calls to retrieve ranked pages from search results as the focus of my project is not on crawling,
   II) Limit the documents to only static htm and html pages,
   III) Limit the number of retrieved documents to 50 or 10 API calls for each query, whichever is met earlier; I can only make a limited number of API calls for a month and wanted to leave you with sufficient balance so you can test it

Here's how the page will look like while the query is under process –



Here's how it's going to look like once it's done and ready with the clustering results –



You can toggle any three clustering results to get a comparative view – (LINGO, KMeans and Suffix Tree)

Example 1: Result of LINGO clustering for query **learn dchinese^**

Example 2: Result of KMeans Clustering for query **blue whale**

Special Note: **I ran into few issues while deploying the website in CIMS servers, often it was memory related, and in other cases it was pretty random. I debugged and fixed all of them I could; due to the lack of resources and control, I could not configure the server the way I wanted for my app, so had to resort for random fixes. It should work most of the times(unless Bing is down), but on rare occasions the website might through an alert message(basically a a generic error message). I can happen due to multiple reasons, so Please try it some other query(possible well formed). You can always try it from console as mentioned in Step A.**

I discussed this in detail with Professor and gave him a demo that the application is in fact working. He was content with it and asked me to make this note to you, so that you are aware of the Website issue.

If you have any questions, feel free to email me: ab5966@nyu.edu

Software/Libraries/Tools used: Carrot2, Apace Lucene, JAVA, Python, CGI, JQuery, AJAX, CSS/HTML