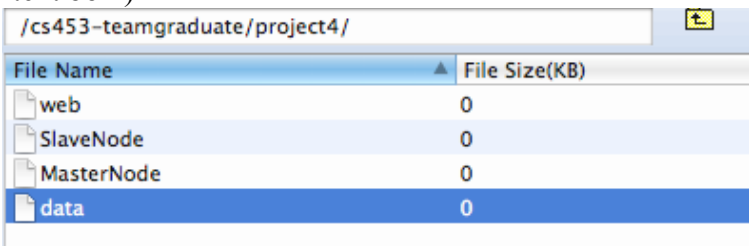
Project 4: Deployment and Usage

I. What you need

1. S3 bucket

You can keep this bucket, it contains all of the data, and source code needed.
Bucket: /cs453-teamgraduate/project4/
(if you can not click the folder *project4* or any folder to view it, you can type the folder name in the text box)



2. AMI

You can search for our AMI in Elasticfox.

a) id : ami-1131b778
b) Maninfest : 536630104582/CS453_teamgraduate_image2

This image contains apache, php, java, s3fs (used to mount s3 bucket from ec2 node)

II. Deployment

1. Slave node

Follow the following steps to start each of 3 slave nodes

a) Launch m1.xlarge instance from the AMI
b) ssh to the instance
c) sudo su
d) mkdir cs453-teamgraduate
e) s3fs cs453-teamgraduate cs453-teamgraduate (mount s3 bucket)
f) cd cs453-teamgraduate/project4/SlaveNode/build
g) ./make (if not work : chmod 755 make)

h) java -client -Xmx1000m SlaveServerThread ***port inverted_list_path***

For slave node 1:

h) java -client -Xmx1000m SlaveServerThread 60000 ../../data/list1
(wait some seconds until you saw: Waiting for incoming connection request…)

For slave node 2:

h) java -client -Xmx1000m SlaveServerThread 60002 ../../data/list2
(wait some seconds until you saw: Waiting for incoming connection request…)

For slave node 3:

h) java -client -Xmx1000m SlaveServerThread 60004 ../../data/list3
(wait some seconds until you saw: Waiting for incoming connection request…)

2. Master node

Follow the following steps to start a master node

a) Launch m1.xlarge instance from the AMI
b) ssh to the instance
c) sudo su
d) mkdir cs453-teamgraduate
e) s3fs cs453-teamgraduate cs453-teamgraduate (mount s3 bucket)
f) cd cs453-teamgraduate/project4/MasterNode/build
g) cp ../../web/.* /var/www*
h) chmod 755 /var/www -R
i) ./make (if not work : chmod 755 make)
j) java -cp .:gson.jar -client -Xmx1000m master/ServerThread ***slave1_ipaddress,slave1_port slave2_ipaddress,slave2_port slave3_ipaddress,slave3_port pagerank_path document_length_path document_title_path document_url_path stop_list_path cached_document_path***

For j), you just need to add the ip addresses of slave nodes into the command below

java -cp .:gson.jar -client -Xmx1000m master/ServerThread ***slave1_ipaddress****,60000* ***slave2_ipaddress****,60002* ***slave3_ipaddress,****60004 ../../data/pagerank.dat ../../data/doclength.txt ../../data/doctitle.txt ../../data/pid_map.dat ../../data/stoplist.txt /home/ubuntu/cs453-teamgraduate/project4/data/*

Note: IP address of a slave can be got by right click on an instance in Elasticfox and select "Copy Private IP Address to the clipboard"

3. Web application

1) Open browser (chrome if you want to see voice search feature)
2) enter : public_dns_master_node/search.html
3) select query type (AND or OR query), and model (TFIDF, BM25, Language model)

Note: Language model is Query likelihood model.

4) and start enter your search query.

III. Usage

1. For suggestion: start showing suggestions after an user types at least 3 characters
2. For "Did you mean":
   a) "Did you mean" algorithm is triggered whenever an user types a space, or press enter in the search box.
   b) "Did you mean" is shown whenever the result from "did you mean" is different from the user's query.
3. For Cached option: some document 's cache is an blank html page because
   a) We zipped, and uploaded 200000 documents (pages.zip) into s3, and unzipped it, some docs could not be unzipped in the cloud. It also took a long time to unzip in the cloud, we could not unzip all 200,000 documents. The total number of successful unzipped documents is about 102,000 documents.
   b) Some document is just blank since we could not download its content in the project 1.