Wikipedia Search Engine

DESCRIPTION

The objective of the mini project is to design a scalable and efficient search engine using the wikipedia data. The search engine should take very little time (less than a sec) to search even the long queries. It supports field queries (for 5 fields- title, infobox, outlinks, category and content) and the index size should be less than 1/4 of the data size. You have to build your own indexing mechanism i.e. you cannot use nutch or lucene toindex the wikipedia data.

The evaluation will be done on 4 parameters - Search time, Search efficiency, Indexing time and Index Size. One is free to use compression techniques, explore several ranking functions (tf,tf-idf, normalized tf, normalized idf etc) and create a secondary index if required.

The language used for the development of search engine can be – C++, JAVA or PYTHON. Using compression techniques can fetch you bonus marks.

Instruction for quantizing.py
You can follow the below steps to split the 39GB (newWiki.xml) xml file into multiple xml files:
1. mkdir Wiki_Split_Files
2. cd Wiki_Split_Files
3. split -b 100M --suffix-length=4 <wiki-xml-dump > (--Make sure xml dump file doesn't exist in the current directory--)
4. cd ..
5. python quantizing.py "Wiki_Split_Files/"
ps: split is a linux command, you can refer man pages for the flag details.

The deadline for uploading first deliverable is 17th Jan where the indexing time and size will be evaluated.
The final evaluation of the systems will take place in 1st week of February.

DELIVERABLES

You need to upload the following deliverables by 17th Jan on the SIEL portal (http://search.iiit.ac.in/courses/www/index.php) :

- tar file of your code.
- Single index file named “Search_Index”.
- a shell script which includes code to compile and run your code.

Copying or even working in Partnership might fetch you “F” Grade.

Enjoy Searching !