HW1

Vegar Engen, Ole Magnus J and Martin Havig

You are asked to plan a web service using Lucene/Solr to answer user questions by searching an NSF abstract database.

Details about dataset and Lucene/Solr could be found at this readme file.

If you plan to use Lucene Java API and Eclipse IDE, this readme file can be useful.

What to do

1. Implement a simple web service to search this dataset.

Check

2. Perform the indexing of 3 parts of this dataset, and estimate the indexing time and disk storage requirement for this database with 1 million records, 10 million records, and 100 million records.

Produced numbers:

What	Time spent	Size	Records
All	0:06:03.020	413.81 MB	132829

Estimate:

Record Size	Indexing time (estimate)*	Size (estimate)**
132 829	0:06:03.020 (actual)	413.81 MB (actual)
1 000 000	0:45:32.987	3115.36 MB
10 000 000	7:35:29.875	31153.59 MB
100 000 000	3 days 3:54:58.752	311535.88 MB

* 0.36589995041 records per ms

** 0.00311535884 mb per record

3. Investigate if you can place index partitions in different machines for these 3 NSF data partitions and serve a query using these machines in parallel. Distribution of search index is discussed here

See HowTo for setup of shards.

What	Time spent	Size
Part1	0:02:36.357	104.04 MB
Part2	0:02:22.725	115.42 MB
Part3	0:01:36.120	151.45 MB

4. Record the average response time for answering a search query from one machine and from multiple machines with distributed index.

Search:

```
(curl
'http://localhost:7574/solr/select?shards=localhost:7574/solr,localhost:7575/solr,localhost:7576/solr&indent=true&q=QUERY')
(curl 'http://localhost:7577/solr/select?indent=true&q=QUERY')
```

Query	**Time (3 Shards)**	**Time (Single)**	**Hits**
martin	160	10	1428
apples	29	7	10
Determinacy	39	2	28
cow	22	8	23
plane	18	9	1237
norway	34	7	134
december	37	13	19980
june+july	103	38	64921
may	26	9	42704
sound+AND+mute	12	13	0

We see that time spent on three shards is significantly higher than with a single one, which is not as expected when the same search produces the same information. This might be cause by the fact that the three shards are run on the same computer on different ports and they are delayed by not optimized parallel computing/not in parallel. Second time something is search for we see that the query time is reduced to close to 1 ms due to caching.

Querys:
May 15, 1991
February 15, 1994
Earth
We saw that the highest results all contained the whole phrase or all parts of it. The ones that contained everything as search for could be longer, so a shorter record with the phrase spread would get a higher score than a longer one.

What to submit

- Turn in the source code directory WITHOUT binary files, using a turnin program (turnin HW1@cs290n directory-name).

What to show during demo (April 29-May 2)

- Show the process for indexing of a sample data file and a query processing for the NSF dataset you have set.
- Show the ranked results for a few queries, explaining why they make sense.
- Explain your performance numbers.
- Explain your finding on distributed processing of a query using multiple machines.

HowTo:

Getting lucene/solr:

1. sudo apt-get install subversion
2. svn checkout http://svn.apache.org/repos/asf/lucene/dev/trunk lucene/dev/trunk
3. sudo apt-get install ant
4. cd /lucene/dev/trunk
5. ant ivy-bootstrap
6. ant
7. cd lucene
8. ant dist
9. cd dist ...

Getting datasett

1. mkdir dataset
2. cd dataset
3. wget http://archive.ics.uci.edu/ml/machine-learning-databases/nsfabs-mld/Part1.zip
4. wget http://archive.ics.uci.edu/ml/machine-learning-databases/nsfabs-mld/Part2.zip
5. wget http://archive.ics.uci.edu/ml/machine-learning-databases/nsfabs-mld/Part3.zip
6. unzip Part1.zip
7. unzip Part2.zip
8. unzip Part3.zip

Setting up shards:

1. cd /lucene/dev/trunk/solr
2. cp -r example example7574
3. cp -r example example7575
4. cp -r example example7576
5. perl -pi -e s/8983/7574/g example7574/etc/jetty.xml example7574/exampledocs/post.sh
6. perl -pi -e s/8983/7575/g example7575/etc/jetty.xml example7575/exampledocs/post.sh
7. perl -pi -e s/8983/7576/g example7576/etc/jetty.xml example7576/exampledocs/post.sh
8. java -server -jar example7574/start.jar
9. java -server -jar example7575/start.jar
10. java -server -jar example7576/start.jar

Adding dataset:

1. java -Dauto -Durl=http://localhost:7574/solr/update -jar /lucene/dev/trunk/solr/example7574/exampledocs/post.jar ./Part1/*/*
2. java -Dauto -Durl=http://localhost:7575/solr/update -jar /lucene/dev/trunk/solr/example7575/exampledocs/post.jar ./Part2/*/*
3. java -Dauto -Durl=http://localhost:7576/solr/update -jar /lucene/dev/trunk/solr/example7576/exampledocs/post.jar ./Part3/*/*
4. All: java -Dauto -Durl=http://localhost:8983/solr/update -jar /lucene/dev/trunk/solr/example/exampledocs/post.jar ./*/*/*