Dorian Karter
March 19, 2013
CS 429 – Information Retrieval
Professor: Dr. Lisa Gandy

# Assignment 2 - Summary

## What I did

I created a java command line app that uses Lucene to build an index from supplied XML file. The app then searches for the strings provided in the assignment description and prints the results.

## How I did it

First I created a class called IndexBuilder. This class builds the index using similar code as was used in the first assignment.

Second step was to create a Searcher class, which provides the interface for querying the index and makes the adjustments to the query in order to accommodate wildcards.

To accommodate wildcards I simply convert query special characters, such as #, to their equivalent Regular Expression. This is done using 'replace' on the string in a method called getRegExQuery.

I then proceeded to create Unit Tests to ensure the correct operation of the Regular Expressions used by the application. The unit tests verify that queries are converted correctly to their equivalent regex, and that the regex used matches/fails against predefined valid/invalid strings.

As was recommended I used RegExpQuery to search the index. The Total Hits and Top 10 Document Identification Numbers are then printed to the output stream.

## What I wish I had known

I wish I had known that the implementation of Regular Expressions in Lucene is different than the one I am used to. For example the \d regular expression to match a single digit does not work with Lucene… I wasted quite a few hours trying to figure out what was wrong and wrote the Unit Tests for that reason, cause I could not figure out what was wrong with my apparently correct code.

Dorian Karter
March 19, 2013
CS 429 – Information Retrieval
Professor: Dr. Lisa Gandy

## Results

==========
Wildcard query: #### generated regex: [0-9][0-9][0-9][0-9]
Total Hits (Documents): 101
Top 10 Documents:
FR941122-0-00001
FR941122-0-00002
FR941122-0-00003
FR941122-0-00004
FR941122-0-00005
FR941122-0-00006
FR941122-0-00007
FR941122-0-00008
FR941122-0-00009
FR941122-0-00010
==========

==========
Wildcard query: environ* generated regex: environ.*?
Total Hits (Documents): 16
Top 10 Documents:
FR941122-0-00030
FR941122-0-00031
FR941122-0-00035
FR941122-0-00036
FR941122-0-00044
FR941122-0-00045
FR941122-0-00049
FR941122-0-00050
FR941122-0-00052
FR941122-0-00054
==========

==========
Wildcard query: agri@ulture generated regex: agri[A-Za-z]ulture
Total Hits (Documents): 8
Top 10 Documents:
FR941122-0-00002
FR941122-0-00004
FR941122-0-00005
FR941122-0-00006
FR941122-0-00053
FR941122-0-00066
FR941122-0-00074
FR941122-0-00091
==========

Dorian Karter
March 19, 2013
CS 429 – Information Retrieval
Professor: Dr. Lisa Gandy

==========
Wildcard query: *ment generated regex: .*?ment
Total Hits (Documents): 92
Top 10 Documents:
FR941122-0-00002
FR941122-0-00003
FR941122-0-00004
FR941122-0-00005
FR941122-0-00006
FR941122-0-00007
FR941122-0-00008
FR941122-0-00009
FR941122-0-00010
FR941122-0-00011
==========

==========
Wildcard query: *ember generated regex: .*?ember
Total Hits (Documents): 53
Top 10 Documents:
FR941122-0-00001
FR941122-0-00002
FR941122-0-00004
FR941122-0-00005
FR941122-0-00007
FR941122-0-00008
FR941122-0-00009
FR941122-0-00010
FR941122-0-00019
FR941122-0-00025
==========

==========
Wildcard query: ####_ generated regex: [0-9][0-9][0-9][0-9][A-Za-z0-9]
Total Hits (Documents): 65
Top 10 Documents:
FR941122-0-00002
FR941122-0-00003
FR941122-0-00004
FR941122-0-00005
FR941122-0-00006
FR941122-0-00007
FR941122-0-00008
FR941122-0-00009
FR941122-0-00010
FR941122-0-00013
==========

Dorian Karter
March 19, 2013
CS 429 – Information Retrieval
Professor: Dr. Lisa Gandy

==========
Wildcard query: # percent generated regex: [0-9] percent
Total Hits (Documents): 0
Top 10 Documents:
==========

==========
Wildcard query: @@@@ generated regex: [A-Za-z][A-Za-z][A-Za-z][A-Za-z]
Total Hits (Documents): 101
Top 10 Documents:
FR941122-0-00001
FR941122-0-00002
FR941122-0-00003
FR941122-0-00004
FR941122-0-00005
FR941122-0-00006
FR941122-0-00007
FR941122-0-00008
FR941122-0-00009
FR941122-0-00010
==========