Report: Project 3 Milestone 2

Joel Fuentes (95686428) and Han Ke (62087924)

February 28, 2015

Continuation of Inverted Index Implementation

As the last milestone reported, the implementation of the index module (inverted index) was built by using a B-Tree data structure stored on disk. The Key value used in this index were the unique *terms* retrieve from the websites in the previous part of the project.
It is important to remark that the B-Tree structure is not built in main memory due to the big amount of memory that the set of website implies. To solve this problem a set of files on disk was used to create a pre-processing with the information needed to build the B-Tree in a second step. These file store all the information needed to create the inverted index by terms and documents (URLs).
Different classes and methods were created for milestone 2. These are:

- *QueryProcessor*: This class has important methods to compute a search based on the query and the indexes.
- *IndexController*: This class has the main method and provides a menu with the following options:
  - Build Index for ics.uci.edu (It may take a long time)
  - Compute Deliverables
  - Execute query

Deliverables

The set of deliverables requested for this part of the project can be obtained by running the program. Results for queries (top 5):

1. mondego

```
mondego: 19 results
1)  http://mondego.ics.uci.edu/
        score: 105.94
        1 words matched
        tf-idf: 5.94
```

```
2)  http://www.ics.uci.edu/community/news/notes/
        score: 105.23
        1 words matched
        tf-idf: 5.23

3)  http://www.ics.uci.edu/~lopes/
        score: 105.23
        1 words matched
        tf-idf: 5.23

4)  http://www.ics.uci.edu/~djp3/classes/2006_03_30_ICS105/Resources/AnteaterIdol.html
        score: 105.23
        1 words matched
        tf-idf: 5.23

5)  http://www.ics.uci.edu/community/news/notes/index
        score: 105.23
        1 words matched
        tf-idf: 5.23
```

2. machine learning

```
machine learning: 2927 results
1)  http://www.ics.uci.edu/~pazzani/Publications/OldPublications.html
        score: 211.48
        2 words matched
        tf-idf: 11.48

2)  http://www.ics.uci.edu/~pazzani/Publications/APubs.html
        score: 211.46
        2 words matched
        tf-idf: 11.46

3)  http://isg.ics.uci.edu/events.html
        score: 209.97
        2 words matched
        tf-idf: 9.97

4)  http://www.ics.uci.edu/~qliu1/MLcrowd_ICML_workshop/
        score: 209.58
        2 words matched
        tf-idf: 9.58

5)  http://www.ics.uci.edu/~qliu1/MLcrowd_ICML_workshop/index.html
        score: 209.58
        2 words matched
        tf-idf: 9.58
```

3. software engineering

```
software engineering: 3241 results
1)  http://www.ics.uci.edu/~wscacchi/Papers/Vintage/Software_Productivity.html
        score: 106.45
        1 words matched
```

```
        tf-idf: 6.45

2)  http://www.ics.uci.edu/~wscacchi/publications.html
        score: 105.96
        1 words matched
        tf-idf: 5.96

3)  http://www.ics.uci.edu/~taylor/Publications.htm
        score: 105.87
        1 words matched
        tf-idf: 5.87

4)  http://www.ics.uci.edu/~wscacchi/
        score: 105.79
        1 words matched
        tf-idf: 5.79

5)  http://www.ics.uci.edu/~andre/publications.html
        score: 105.73
        1 words matched
        tf-idf: 5.73
```

4. security

```
security: 1005 results
1)  http://drzaius.ics.uci.edu/~swirl/impromptu-0.30/apidocs/index-all.html
        score: 106.84
        1 words matched
        tf-idf: 6.84

2)  http://www.ics.uci.edu/~gts/pubs.html
        score: 106.72
        1 words matched
        tf-idf: 6.72

3)  http://drzaius.ics.uci.edu/~swirl/impromptu-0.20/apidocs/index-all.html
        score: 106.38
        1 words matched
        tf-idf: 6.38

4)  http://drzaius.ics.uci.edu/~swirl/impromptu-0.30/xref/edu/uci/isr/impromptu/repository/ProxyF
        score: 105.62
        1 words matched
        tf-idf: 5.62

5)  http://drzaius.ics.uci.edu/~swirl/impromptu-0.30/xref/edu/uci/isr/impromptu/security/LevelHan
        score: 105.58
        1 words matched
        tf-idf: 5.58
```

5. student affairs

```
student affairs: 8849 results
```

```
1)  http://www.ics.uci.edu/ugrad/qa/
        score: 211.08
        2 words matched
        tf-idf: 11.08

2)  http://www.ics.uci.edu/ugrad/qa/index.php
        score: 211.08
        2 words matched
        tf-idf: 11.08

3)  http://www.ics.uci.edu/ugrad/
        score: 208.99
        2 words matched
        tf-idf: 8.99

4)  http://www.ics.uci.edu/ugrad/index.php
        score: 208.99
        2 words matched
        tf-idf: 8.99

5)  http://www.ics.uci.edu/grad/index.php
        score: 208.81
        2 words matched
        tf-idf: 8.81
```

6. Crista Lopes

```
Crista Lopes: 161 results
1)  http://www.ics.uci.edu/community/news/notes/notes_2007.php
        score: 213.58
        2 words matched
        tf-idf: 13.58

2)  http://vcp.ics.uci.edu/content/dvas
        score: 212.46
        2 words matched
        tf-idf: 12.46

3)  http://www.ics.uci.edu/community/news/notes/
        score: 211.12
        2 words matched
        tf-idf: 11.12

4)  http://www.ics.uci.edu/community/news/notes/index.php
        score: 211.12
        2 words matched
        tf-idf: 11.12

5)  http://www.ics.uci.edu/community/news/notes/index
        score: 211.12
        2 words matched
        tf-idf: 11.12
```

7. REST

```
REST: 599 results
1)  http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
        score: 107.23
        1 words matched
        tf-idf: 7.23

2)  http://www.ics.uci.edu/~fielding/pubs/dissertation/evaluation.htm
        score: 106.82
        1 words matched
        tf-idf: 6.82

3)  http://www.ics.uci.edu/~kay/courses/i41/answers.html
        score: 106.59
        1 words matched
        tf-idf: 6.59

4)  http://www.ics.uci.edu/~kay/courses/141/schemenotes.html
        score: 105.9
        1 words matched
        tf-idf: 5.9

5)  http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart
        score: 105.9
        1 words matched
        tf-idf: 5.9
```

8. computer games

```
computer games: 22743 results
1)  http://www.ics.uci.edu/~magda/cs620/announceOG.html
        score: 208.72
        2 words matched
        tf-idf: 8.72

2)  http://www.ics.uci.edu/~wscacchi/
        score: 208.61
        2 words matched
        tf-idf: 8.61

3)  http://sli.ics.uci.edu/Classes/2009W-Comments
        score: 208.13
        2 words matched
        tf-idf: 8.13

4)  http://www.ics.uci.edu/~eppstein/cgt/bib.html
        score: 208.03
        2 words matched
        tf-idf: 8.03

5)  http://www.ics.uci.edu/~eppstein/cgt/
        score: 207.86
        2 words matched
        tf-idf: 7.86
```

9. information retrieval

```
information retrieval: 24018 results
1)  http://www-db.ics.uci.edu/pages/research/mars.shtml
        score: 208.77
        2 words matched
        tf-idf: 8.77

2)  http://www-db.ics.uci.edu/pages/research/mars/index.shtml
        score: 208.77
        2 words matched
        tf-idf: 8.77

3)  http://www-db.ics.uci.edu/pages/research/mars/
        score: 208.77
        2 words matched
        tf-idf: 8.77

4)  http://www.ics.uci.edu/~gbowker/converge.html
        score: 208.32
        2 words matched
        tf-idf: 8.32

5)  http://www.ics.uci.edu/~kobsa/privacy/German.htm
        score: 208.2
        2 words matched
        tf-idf: 8.2
```