CSE474 - Information Retrieval and Extraction

Mini-Project (Phase 1)

Search engine for Wikipedia

Note:

- This article highlights the deliverables for the first phase of the mini-project.
- The hard deadline for the first phase is **Jan 21st 19:30:00 IST**. [Please submit the deliverables well in advance!]
- Usage of libraries like Lucene, WikiXMLj is strictly prohibited. [If we allow, you will miss the fun in coding a search engine from scratch.]
- Tutorials will be conducted very soon. [Don't wait for the tutorial to happen. Please start coding NOW itself.]
- TA office timings, submission instructions and phase 2 specifications will be announced soon.
- Courses mailing list: http://lists.iiit.ac.in/mailman/listinfo/cse474
- Please post your queries in the http://moodle.iiit.ac.in.

Task: Construct the Inverted Index from the given small snapshot of Wikipedia dump.

Basic Stages (in order):

- XML parsing [Prefer SAX parser over DOM parser. If you use DOM parser, you can't scale it up for the full Wikipedia dump later on.]
- Tokenization
- Case folding
- Stop words removal
- Stemming [Recommended Stemmer: http://tartarus.org/~martin/PorterStemmer/]
- Posting List / Inverted Index Creation
- Optimize

Desirable Features:

- *Support for Field Queries.* Fields include Title, Infobox, Body, Category, Links, and References of a Wikipedia page. This helps when a user is interested in searching for the movie 'Up' where he would like to see the page containing the word 'Up' in the title and the word 'Pixar' in the Infobox. You can store field type along with the word when you index.
- *Index size should be less than ¼ of dump size.* [You can experiment with different index compressing techniques.]

- *Scalable index construction* [See Chapter 4 in the 'Intro to IR' book.]

Reference:

- *Readings:*
  - http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf [Chapters 1-5]
  - Information Retrieval: Algorithms and Heuristics. D.A. Grossman, O. Frieder. Springer, 2004.
- *Videos:*
  - https://class.coursera.org/nlp/lecture/178
  - https://class.coursera.org/nlp/lecture/179
  - https://class.coursera.org/nlp/lecture/180

Dataset Link for IIIT residents (~100 MB):

- http://10.2.4.50:8080/ire-wiki-search.tar.gz (or)
- http://10.2.4.182:8080/ire-wiki-search.tar.gz

Dataset Link for others (~100 MB):

- https://www.dropbox.com/s/0ymmsga6cawgxe1/ire-wiki-search.tar.gz?dl=0