CS 4731 - 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 Aug 23 23:55:00 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.]

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
- 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 1⁄4 of dump size. [You can experiment with different index compressing techniques.]
- Scalable index construction [See Chapter 4 in the 'Intro to IR' book.]

Evaluation Criteria:

Evaluation for phase I will be on below two criteria:

(a) Index creation time: less than 60 secs for Java, CPP and for python it's less than 150 secs.
(b) Inverted index size: expected size is 25-30 mb.

Reference:

Readings:
o http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf [Chapters 1-5]
o Information Retrieval: Algorithms and Heuristics. D.A. Grossman, O. Frieder. Springer, 2004.

Videos:
o https://class.coursera.org/nlp/lecture/178
o https://class.coursera.org/nlp/lecture/179
o https://class.coursera.org/nlp/lecture/180