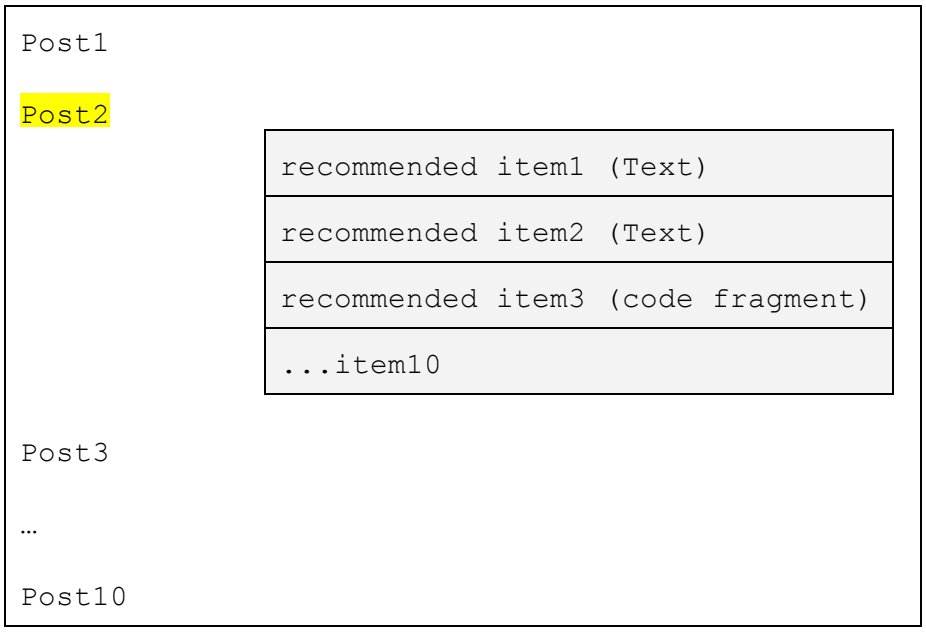
Assignment #2: (Content-based) Recommendation

Tasks:
In this assignment, you will implement a content-based recommender via an web app by recommending similarity-based Java programming wikibooks content (https://en.wikibooks.org/wiki/Java_Programming) to Stackoverflow.com data set (10 posts ONLY).

1. crawl java programming wikibooks pages
2. use Apache Lucene (and/or SOLR) or any Lucene compatibles to index crawled content
3. build a simple web app to display the provided StackOverflow 10 posts, by selecting each post, list the top 10 relevant wikibooks *items* (from your indexed documents).

For instance of the browser outlook:



Data & demo codes are ready for download on Blackboard. Data sets are provided by the instructor for educational demonstration purpose only. You are prohibited to distribute or to share. You are free to edit and transform the sample codes.

Evaluation: (105/100)

- *Content Collection (20pt):* {-20: no Java Programming wikibooks content; -5~-10: not enough content(less than 50% of the wikibooks pages); bonus(+3): include Oracle The Java™ Tutorials}
- *Content Indexing (50pt):* {-50: hardcoded post<->wikibooks content; -10: index by pages; +2 stemming}
- *Web app (20pt):* Display data & recommendations and how you index the content on web pages {-20: broken codes; -10: no explanations on how you index the content}
- *Originality (10pt):* Use your imagination and creativity to implement ANYTHING (can be UI, can be customized scoring methods) outside the box but related to this assignment. Describe the

method explicitly on the bottom of the page. i.e. Lucene supports different ranking system based on similarity with API (BM25, VSM, even language models)

Submission:

1. This is an individual assignment.
2. Your submission should be executable offline in my local machine. You must enclose a zip file for all the source code and readme.txt if the entry page is not **index.html.**
3. File must be named as Assign2-YourFirstnameLastname (i.e. Assign2-SharonHsiao)
4. Submit it through Blackboard.
5. Deadline: **Wednesday Feb.24/2016, by noon.**

Your assignment will be discussed in class or on blog.

Recommended reading:

download Lucene here: http://lucene.apache.org
API: http://lucene.apache.org/core/5_4_1/core/overview-summary.html#overview_description
http://www.lucenetutorial.com/index.html
http://www.lucenetutorial.com/lucene-query-syntax.html

Some other suggested tools to explore:
(1) Content-based: Apache Lucene/SOLR, PyLucene
(2) Collaborative Filtering based: LensKit (http://lenskit.org)
(3) Case-based: jCOLIBRI (https://gaia.fdi.ucm.es/)