Birla Institute of Technology & Science – Pilani
Second Semester 2014-15

Date: 04.04.2015

Take Home Assignment- 2

Course Name	**: Information Retrieval**
Course No	**: CS F469**
Submission Deadline	**: 10:00 AM On 06.04.2015**
Max Marks	**: 5M**
Type	**: Individual Assignment**

NOTE 1: **You are free to implement in any language of your choice. You should be comfortable in explaining your approach.**
NOTE 2: **Assignments of only those students will be evaluated who have attended lab on 04.04.2015 (Saturday). It is also important to note that the deadline is a hard deadline.**
NOTE 3: **Developing a GUI is good but it is not mandatory to do so. You may use existing IR_tool for your experiments.**
NOTE 4: **Clearly State your assumptions, if any.**
NOTE 5: **Solutions have to be mailed to (h2009399@pilani.bits-pilani.ac.in CC: mehala@pilani.bits-pilani.ac.in) with subject "Information Retrieval Take Home Assignment 2". In Mail Content: Add your ID No. and the timing of your Evaluation slot during 08th April-10th April 2015.**

AIM: The purpose of this task is to implement a simple Meta-Search engine system and its ranking and evaluation system.

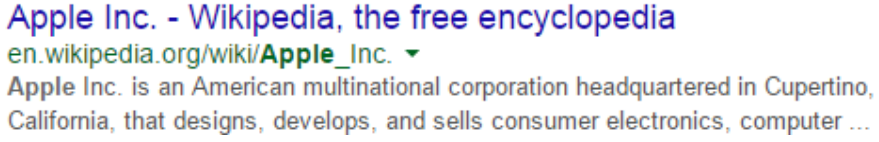
BASIC:
A meta search engine is a search tool that uses other search engines' data to produce their own results from the Internet. Metasearch engine takes input from a user and simultaneously send out queries to multiple search engines and aggregates results from multiple search engines. These aggregated documents are ranked and presented to the users.
In this assignment, we will implement a simple meta search engine which produces aggregated documents which are extracted from Google and Yahoo! search engine. In the literature, lot of meta search aggregated documents ranking mechanism have been proposed and compared. In this work, we will apply a simple ranking method. The purpose of this work is to understand the working principle of meta search engine rather than focusing on effective meta search engine.

PROBLEM:
Step 1: Document Extraction from a search engine

……

Assumptions:

- In this assignment, assume document content as only its summary/snippet and the title for simplicity. It is to note clearly that based on our assumption, document does not refer whole content. Thus, you need not retrieve full document content (say complete Wikipedia article on Jaguar) for this assignment. Only extract Title and Snippet as a document content. ***[Note: This assumption is not true for the real meta search engine. This assumption is to simplify the work.]***
  The following figure shows the title and snippet regions of a search result.



<----- Title

<----- Snippet

- Ignore **Image/Video search results.**

……..

Extract top-n (n=30) documents [after ignoring image/video files] related to the query *"Jaguar"* from two search engines namely *Google* and *Yahoo!* and store the retrieved top-n results in the files "Jaguar_Google.txt" and "Jaguar_Yahoo.txt" respectively. ***[Note: These files should contain 30 lines with each line representing one document i.e. each document will be represented as unique_docID followed by a tab space followed by (Title+Snippet)].***

Step 2: Aggregated Document Collection

Combine the results obtained from *Google* and *Yahoo!*. In this task you are required to consider only unique documents *(Note: State the assumption if any)* from the extracted set. Store this in a file ***UniqueDocuments.txt. Note: Here also, each document will be represented as docID followed by a tab space followed by (Title+Snippet).***

Step 3: Ranking

a) Approach 1: (Best Rank approach)

This approach, place a URL at the best rank it gets in any of the search engine rankings.
That is,
MetaRank (x) = Min(GoogleRank(x),YahooRank(x))
// MetaRank refers rank assigned by meta Search engine

Clashes are avoided by an ordering of the search engines based on popularity. That means, if two results claim the same position in the meta-rank list, the result from a Google search engine is preferred to the result from Yahoo! search engine.

Store the results obtained based on this approach in a file **ResultantRanks_A1.txt**

b) Approach 2: (Borda's Positional approach)

In this approach, MetaRank of a url is obtained by computing the Lp-Norm of the ranks in different search engines.
That is,
MetaRank(x)= $[\Sigma$ (GoogleRank (x) ^{p}, YahooRank $(x)^{p})]^{1/p}$
// MetaRank refers rank assigned by meta Search engine
In this approach, consider the L1-Norm which is the sum of all the ranks in different search engine result lists. Clashes are avoided by an ordering of the search engines based on popularity. That means, if two results claim the same position in the meta-rank list, the result from a Google search engine is preferred to the result from Yahoo! search engine.
Store the results obtained based on this approach in a file **ResultantRanks_A2.txt**

Step 4: Evaluation

Evaluating Performance is very important for any system. We will evaluate our system based on two factors: retrieval accuracy and ranking accuracy.

- Manually assign ranking to the above document set (UniqueDocuments.txt) with the intention "jaguar cars". Store this in a file **RankedDocuments.txt**. This file is the ground truth file to be used for evaluation.
- ***Retrieval Accuracy***

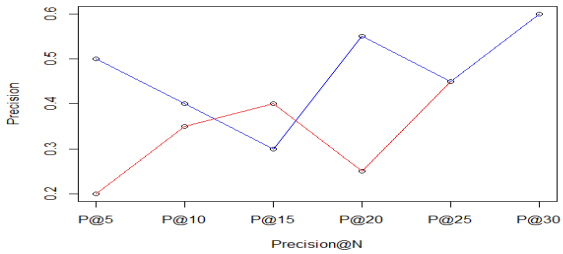
  Note: To simplify manual labeling of document as relevant or irrelevant, label the document as relevant if the document appears in top N results from RankedDocuments.txt. Otherwise, label the document as irrelevant. For Example**:** while calculating Precision@5, Top 5 Results from ***RankedDocuments.txt*** will be considered as relevant. Others will be considered as irrelevant.

1. Find precision@5, precision@10, precision@15, precision@20, precision@25, precision@30 with respect to the ranked aggregated collections formed based on ranking approach 1 and approach 2 (i.e. ResultantRanks_A1.txt and ResultantRanks_A2.txt respectively).

2. Plot the obtained results as follows:
   X-axis *needs to have labels such as P@5,P@10, P@15,P@20, P@25,P@30.*
   Y-axis *needs to have values obtained for P@5,P@10, P@15,P@20, P@25,P@30. i.e. will vary between 0 & 1 (Both Inclusive)*

Plot needs to represent different (2 curves) one for each approach. (***Note:*** You may plot curve using Excel or R or any plotting software.)

A sample plot (**with random values**) is being provided below *(Note: This is only for explanatory purposes and has no relation with actual results).*



3. Calculate MAP for both the approaches:
   MAP=(P@5+P@10+ P@15+P@20+ P@25+P@30)/6

- **Ranking Accuracy**

  In this task we need to use Spearman Correlation Coefficient (based on ρ $(refer\ eq.1)$) to evaluate ranking effectiveness of both the approaches in step3 in conjunction with **RankedDocuments.txt**.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (1)$$

  Where ρ (say rho) is Spearman Correlation Coefficient.

  To know more about Spearman Correlation Coefficient you need to go through following link:
  http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

Deliverables:

1. Working solution of the problem in any language of your choice.
2. Help file briefing your setup, i.e. input requirements, class descriptions, function descriptions, name of the output file generated
3. Approach file: A small document describing the approach you used to develop your system. Keep it short and simple, verbose documents will not yield extra marks. You may include a system diagram (not mandatory, but for your own understanding) to explain your system flow.

----------**BEST OF LUCK**----------