Birla Institute of Technology & Science – Pilani
DSE 2018 Batch, Second Semester 2019-2020 - Second Semester

Take Home Assignment

Course Name : Information Retrieval
Course No : DSECL ZG537
Submission Deadline : 5th JAN 2021
TYPE : Group Assignment

NOTE 1: **You are free to implement in any programming language of your choice. You should be comfortable in explaining your approach.**
NOTE 2: **Developing a GUI is good but it is not mandatory to do so.**
NOTE 3: **Clearly State your assumptions, if any.**
NOTE4: **Deliverables/Solutions have to be mailed to (anand.a.bhosle@wilp.bits-pilani.ac.in) with subject "Information Retrieval Assignment". In Mail Content: Add your Group No.**

AIM:

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

II. Literature survey: Focusing on the techniques for building highly scalable and effective metasearch engines and related techniques. ***Note:*** You can pick a topic of your own for the literature survey/case study related to the topic "building highly scalable and effective metasearch engines and related techniques". Send me the chosen topic (with few lines of description about the topic) for the literature survey by mail **(anand.a.bhosle@wilp.bits-pilani.ac.in)** ***on or before 23rd Dec, 2020. Please note that the topics for the literature survey will be accepted as per the FIFO basis.***

AIM: I

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 (i.e. Query) 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 news which are extracted from ***any two news search engine*** like Google News and Yahoo! News search engine. The purpose of this work is to understand the working principle of meta search engine rather than focusing on effective meta search engine.

PROBLEM TASKS:

Task 1: News Extraction from a search engines

…………………………………………………………………………………………………
Assumptions:

- In this assignment, assume document content as only its summary/snippet, the title, Category and date&time of the news 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 article on a news) for this assignment. Only extract summary/snippet, the title, Category and date&time of the news as a document content. ***[Note1: This assumption may not be true for the real meta search engine. This assumption is to simplify the work.]***
- ***Note2:*** Ignore **Image/Video search results.**
- ***Note3:*** Choose any two category for the same purpose and choose any random date for extracting news for this task.

- ***<u>Propose a method suitable for processing and storing the extracted information from different search engines</u>***

…………………………………………………………………………………………..

<u>Task</u> 2: Aggregated News Collection

Combine the news information (with respect to a chosen date) obtained from *any two news search engine.*

- In this task you are required to consider only unique news documents *(Note: State the assumption/method if applicable)* from the extracted set.
- ***<u>Propose a method suitable for processing and storing the unique documents information extracted from different news search engines</u>***

<u>Task</u> 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(Rankof searchengine1(x),Rankof searchengine2(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 any one search engine is preferred to the result from another search engine.
<u>Note1: Indicate your assumption regarding the Popularity of the News Search engines.</u>

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

b) Approach 2:

Propose your own method for ranking the news documents from the **Aggregated News Collection (Taks2)**
Write the results in the document named ResultantRanks_A2.txt

Note: ***ResultantRanks_A2.txt*** should contain atleast document name, its score and rank.
State the other valuable information to be recorded in this file according to your understanding. Justify the same.

<u>Task</u> 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 [from the **Aggregated News Collection (Taks2)]**

- Store this in a file **RankedDocuments.txt**. This file is the ground truth file to be used for evaluation.
- ***Retrieval Accuracy***

  <u>Note:</u> 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

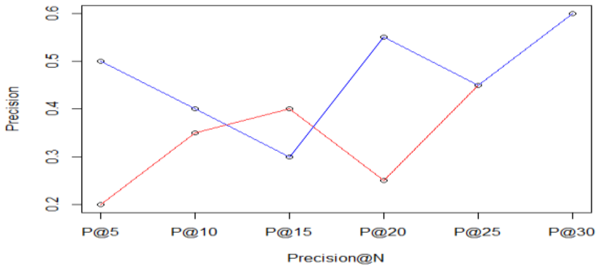
ranking approach 1 and approach 2 (i.e. ResultantRanks_A1.txt and ResultantRanks_A2.txt respectively).

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/package.)

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



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 etc.
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 to explain your system flow.

AIM: II

Instruction(s)

You can pick a topic of your own for the literature survey/case study related to the topic “building highly scalable and effective metasearch engines and related techniques”. Send me the chosen topic (with few lines of description about the topic) by mail (anand.a.bhosle@wilp.bits-pilani.ac.in) ***on or before 23rd Dec 2020.***

Note: Topics will be accepted as per the FIFO basis.

General steps

- Pick a topic
- Survey related work
- Write a report
- Submit the report

Deliverables:

Report Submission

Table of Contents

Contents

i. Title page
ii. Acknowledgement
iii. Undertaking
iv. List of Symbols & Abbreviations used
v. Abstract
vi. Table of contents
vii. Contents
1. Introduction
2. Related work
3. Literature survey: Description of the existing methods and limitations
4. Literature survey: Precise description of experiment setup, tools and experimental data set
5. Thorough discussion on the existing methods and detailed analysis of the experimental results of the existing methods

viii. Conclusion and Future Scope
ix. Appendices
x. Bibliography/References
xi. List of Publications/Conference Presentations, if any.

<u>Note: Maintain the uniformity of formats of the references</u>

Illustrative Examples of Citation of References:

1. Book: A. Gelb, Applied Optimal Estimation. Cambridge, M.A.; M.I.T. Press, 1974
2. A paper in Conference or Symposium Proceedings edited Published by Book Company:
R.E. Kalman, `New Methods in Wiener filtering theory', in Proc. First Symposium on Engineering Applications of Random Function Theory and Probability' J.L. Bogdanoff and F. Kozin, Eds. New York, Wiley, 1963, pp. 270-388
3. A Journal Paper:
R.E. Kalman and N.S. Pucy, `New results in linear filtering and prediction theory', Trans. ASME, J.Basic Eng., Vol. 83-D, pp. 95-108, Mar. 1961
4. A Conference Paper: M. Vidyasagar and N.K. Bose, `Input-output stability of linear systems defined over measure spaces', in Proc. Midwest Symp. Ciro, Syst., Montreal, P.O. Canada, Aug. 1975, pp 394-397
5. A Ph.D. dissertation or Dissertation
A.C.G Viera, `Matrix, orthogonal polynomials, with applications to autoregressive modeling and ladder forms', Ph.D. Dissertation, Stanford Univ., Stanford, CA, Dec. 77

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