BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
HYDERABAD CAMPUS
CS F469 - Information Retrieval

Assignment – 1
First Semester 2016-17

TOTAL MARKS: 40 Marks
DUE DATE: 24/10/2016

This assignment is aimed at designing and developing one's own text based information retrieval system. The assignment is divided into two phases. Phase-1 focuses on building the indexing module that takes in a large collection of data as input, and produces a searchable and persistent data structure. Phase-2 of the assignment aims at building the searching module according to Boolean, Vector Space or Probabilistic model of Information Retrieval.

The assignment can be done in groups of at most 4 (FOUR) members. All the group members are expected to contribute to all the aspects of the assignment namely, design, implementation, documentation and testing.

Tips and Tricks:

- Phase-1 comprises of two major components – Tokenization and Normalization.
- Vector Space Model and Probabilistic Model require additional information like number of documents, maximum term frequency, length of the document vector etc. Ensure that this information is stored in well-defined data structures for easy access.

Corpus:

Any corpus can be used for the assignment. You may also choose a corpus from the following:

1. SNAP – Online Communities
2. SNAP – Online Reviews
3. SNAP – Wikipedia Articles
4. SNAP – Wikipedia
5. Arizona State University - TUAV
6. NLTK Data

Programming Languages:

The assignment can be implemented in any programming language of your choice.
STL's and inbuilt packages can be used only for Normalization (C++'s Boost Library, Python's NLTK Package etc.). You are expected to code the core functionality of the model that you choose (TF-IDF in case of Vector Space model etc.)

Phase-1 Information:

Tokenization:
For this step you can use any standard tokenizer or inbuilt package. Following are a few sources:

- Python's NLTK package.
- Stanford Tokenizer.
- TM package of R.

Stemming:
Martin Porter's 'Porter Stemmer' can be used for this purpose. Implementation in multiple languages can be found in the above link.

Phase-2 Information:

Querying and Searching:
The querying module should accept queries from the user and search for the documents using the data structures produced during Phase-1. Implement the search using Boolean, Vector Space or Probabilistic model to rank the documents by carefully choosing the weighting function.
The search interface should contain the following:

- Provision to input the search query.
- View the list of the search results. Each result should display the Document ID, URL or the Title of the document.

Deliverables:

The final submission must contain the following documents:

1. **Design Document** – This document should contain the description of the application's architecture along with the major data structures used in the project. Precision and Recall, if possible, should also be calculated.
2. **Code** – The code should be well commented.
3. **Documentation** – All the classes, functions and modules of the code must be documented. Software that automatically generate such documents can be used – pydoc for Python, Eclipse for Java etc.
4. **README** – The README file should describe the procedure to compile and run your code for various datasets.

Submission Guidelines:

All the deliverables must be zipped and submitted to **bphc.ir@gmail.com** latest by **September 24, 2016.**
You are expected to demo your application and present your results as per the schedule that will be made available.

Evaluation Criteria:

S.No.	***Task***	***Marks***
1	Tokenizing and Normalizing	5
2	Efficient Dictionary Construction	10
3	Index Construction	5
4	Accurate Data Retrieval	5
5	Viva	5
6	Novelty / Out-of-the-box thinking (Anything that is not covered in the lectures.)	10
	TOTAL	40