Information Retrieval and Extraction [CSE454]

Assignment 1

Part 1: Visible and Invisible Web

1.1 Define Visible Web and Invisible Web
1.2 Discuss atleast three ways to estimate size of the invisible web
1.3 What is the need to measure the size and extent of invisible web

Mention references to your answers. Part 1 of the assignment should be uploaded on the Courses Portal in doc or pdf format. Name the file as <RollNo>_a1.pdf

Part 2: Understanding Indexing Process

Take five documents (eg. news articles, informational text, literary text) written in your mother-tongue. The steps to be followed are :-

i. Case Folding (if required, as in english)
ii. Stop Words Removal : remove most common words in the language that don't contribute to the bag of word representation of text, eg. articles, pronouns, conjunctions etc.
iii. Stemming : find root word, e.g. Beginning : begin
iv. Term Frequency and Indexing
v. Create Posting List.

An example for English is as follows :-

Document #1
Source : Lord of The Rings, Fellowship of the Ring

One Ring to rule them all, One Ring to find them,
One Ring to bring them all and in the darkness bind them.
In the Land of Mordor where the Shadows lie.

Word Count :

one	3	bring	1	lie	1
ring	3	dark*	1	mordor	1
all	2	find	1	rule	1
bind	1	land	1	shadow*	1

We have converted all the letters to lowercase (will not be required for hindi or telugu), and removed the stop words like 'to', 'them' etc. Then darkness and shadows are stemmed. If there is one more line that says “Dark Lord wanted the ring”, count for 'dark' will become 2, counting for 'darkness' as well as 'dark'.

You should write both text as well as the final word count. Make sure that each text contains atleast 80 words. You need to work on 5 such texts. Texts in English will **not** be accepted. Remember to mention the source.

Once you get details for five documents, you need to recompile them to create final posting list. Let us assume that we have another documents that may or may not contain these words. Posting list will contain all the words we discovered. Say, our Document # 2 talks about Harry Potter and

mentions the **dark lord** as well. So the posting for word dark becomes:-

dark – doc1:1, doc2:3

This line implies that the word dark is present in document 1 one time, and document 2 three times. Such a list should be created for each word whose count has been known.

This assignment will build your intuition for Phase 1 of the Miniproject.

Part 2 is to be **handwritten** and handed over to the TAs in the tutorial on Friday. If you are unable to attend the tutorial, submit your assignment in the IE Lab, Vindhya B2 – 102 before the deadline.

Deadline : Friday, 9pm

Please note all deadlines are hard deadlines. Failing to submit on time will lead to zero marks for the assignment. No requests for extension will be entertained.