Information Retrieval Practical Assignment No 1

Roll nos: 1 , 57, 96

1. Choose any corpus available on the internet freely. For the corpus, for each document, count how many times each stop word occurs and find out which are the most frequently occurring stop words.
   Further, calculate the term frequency and inverse document frequency as (Log of no of documents / no of documents having the term). The motivation behind this is basically to find out how important a document is to a given query.
   For e.g.: If the query is say: "The brown crow". "The" is less important. "Brown" and "crow" are relatively more important. Since "the" is a more common word, its tf will be high. Hence we multiply it by idf, by knowing how common it is to reduce its weight.

Information Retrieval Practical Assignment

Roll nos: 2, 58, gopal zalke

2. Implement cosine similarity ranking on a corpus. Choose any corpus available on the internet freely.

Information Retrieval Practical Assignment

Roll nos: 3, 63

3. Choose any corpus available on the internet freely. Not necessary to create an inverted index. Just generate the vocabulary. Download and run Porter Stemmer. Execute the stemmer over terms in the vocabulary to create sets of equivalent terms, all of which stem to the same root form. Which set is largest? Identify a few sets that are inappropriately conflated by the stemmer.

Information Retrieval Practical Assignment

Roll nos: 15, 66

4. Implement an algorithm for intersection of posting lists in a positional index.

Information Retrieval Practical Assignment

Roll nos: 9, 68

5. Implement an algorithm for intersection of posting lists using skip pointers.

Information Retrieval Practical Assignment

Roll nos: 26, 69, gaurav krishna

6. Design and implement a Boolean retrieval system with a suitable interface for it too.

Information Retrieval Practical Assignment

Roll nos: 36, 72

7. Implement tolerant retrieval scheme using edit distance algorithm.

Information Retrieval Practical Assignment

Roll nos: 34, 49, 84

8. Demonstrate a simple program which simulates how a precision recall curve follows a saw tooth curve.

Information Retrieval Practical Assignment

Roll nos: 38, 45, 53

9. Modify edit distance algorithm in the following way:
   Different weights for different operations
   a. Replacement costs depend on proximity of characters in keyboard
   b. Takes care of common keyboard errors
      E.g. Cost of replacing s by a is less than replacing s by t

Information Retrieval Practical Assignment

Roll nos: 40, 52

10. Suppose that we are designing a program to simulate the search in a dictionary. Words appear with different frequencies, however and it may be the case that a frequently used word which is in the stop list like “the” appear far from the root if they are sorted lexicographically while a rarely used word such as consciousness appears near the root. We want that the words that occur frequently in the text to be placed nearer to the root. Moreover, there may be words in the dictionary for which there is no definition. Organize an optimal binary search tree that simulates the storage and search of words in a dictionary.