Information Retrieval and Web Analytics

Semester 2 – 2020.

Assignment 1 – 10%

This is an Individual Assignment

Due Date – 15th August, 2020

Marks will be reduced if code is copied from Internet or from a fellow student.

More marks will be given to the originality of the code.

The written programs should be tested by proper inputs. The input and the output of the program should be clearly shown in the code.

Write the python program for the following Information Retrieval Techniques. You can utilize the given corpus to test your program.

Q1). (25 Marks)

Let us have the following collection of documents.

D1: “Frodo stabbed the orc with the red sword”.

D2: “Frodo and Sam used the blue lamp to locate orcs”.

D3: “Sam killed many orcs in Mordor with the blue sword”.

a). Write a program to identify the terms in the document. (5 marks)

b). Build a term document incidence matrix (5 Marks)

c). Using the above matrix identify the documents that are relevant for the following queries?

q1: (Frodo AND orc AND sword) (5 Marks)

q2: (Sam AND blue AND NOT Frodo) (10 Marks)

Q2). (45 Marks)

Use the documents in the Reuters training data set provided by NLTK for this question. Import nltk library and download the files (nltk.download()). The documents will be located at ../nltk\_data/corpora/reuters/training. In order to collect the vocabulary, you need to apply tokenization and stemming on the document text.

1. Build modules to tokenize words and sentences. (10 Marks)
2. Build module for Porter Stemmer to do stemming. (10 marks)
3. Build module to apply case folding i.e. convert words to lower case. (5 Marks)
4. Build inverted index and posting list as python dictionary. (10 Marks)
5. Build a generalized module to merge any number of posting lists (10 Marks)

Q3). (30 Marks)

Build modules to perform the following IR techniques. Test your modules using proper strings.

1. Create Permuterm Index for any given word. (10 Marks)
2. Find out the Levenshtein distance between two words. (10 Marks)
3. Implement the SOUNDEX algorithm. (10 Marks)