Analysis

Information Retrieval
Assignment 2

Submitted by:
Himanshu Aggarwal
MT17015

TF-IDF Score

The tf-idf score for a term in corresponding to a document has been computed using the formula:

tf-idf = tf X idf

where

tf = log(frequency) + 1 , if frequency>0
0 , otherwise

idf = log(number of document / df) + 1

where, df = number of documents where the term appears atlas once

Cosine Score

Vectors of the tf-idf values for query terms, as well as for each document has been formed and used for representing them in the vector space. These vectors are then used for calculating the cosine similarity between the query term and the documents.

Spelling Correction

This module has been implemented using autocorrect package in python. Has been implemented on the query terms.

Numerical Queries

This model has been implemented using word2number and num2word packages in python. Whenever a query contains number, its word equivalent term is added to the input query. Similarly, if a query term in a number in words, it is converted to number form, and added to the input query.

Attention to Title of the documents

The titles of the documents have been fetched from the index.html file and the terms occurring in the title are then given a higher weight in tf-idf value corresponding to that particular document.

Cache

For this, simply the queries and the retrieved results are being stored in a JSON file. The terms are removed from the cache according to the sequence in which they arrived. A total of 20 cache entries are being maintained.