**CSCE 5200 Information Retrieval and Web Search**

**Project 3: Query Retrieval and Performance Analysis**

**Project Description**

The words **query** and **topic** are interchangeable.

For this project, you are required to build the Query Processor. You will use the same dataset as in the previous two projects.

In this project, you will need to implement the query processing and retrieval portion of the search engine (built on your Project 2 implementation). Your code should support the Vector Space Model and use Cosine Similarity as relevance measure. Your IR Engine should be capable of calculating **TF\*IDF weights** for all the terms in the collection and in the query.

(Note: You only need to store term frequency (tf) in the forward and Inverted Index (actually this is what we did in project 2). The IDF and cosine similarity measure can be computed at runtime.)

The Vector Space Model, cosine similarity measure are detailed in chapter 6 and class slides. Figure 6.14 describes the basic algorithm for computing vector space scores using inverted index.

There is a query file **topics.txt** containing four queries. You need to process them, search for each query in your index, rank the documents retrieved and store the output in a file. The format in which you have to store the output is explained in the readme.txt file.

Each query in the file contains additional information which you can make use of for this task. In particular, each query has three fields – title, description and narrative. For this task, you can make a comparison of performance when you consider only the main query (title), when you consider the description along with the main query (description + title), and when you consider the narrative along with the main query (narrative + title). For performance measures, you can use Precision and Recall introduced in class.

**Resources to be provided**

Following are the files that you will need for this project:

* main.qrels – Relevance judgments file
* topics.txt – Queries
* sample\_output.txt – A sample file showing how the output of your Processor should look like.
* readme.txt – Explains the format of each file in the directory.

**What to submit**

* Source code via online submission
* vsm\_output.txt file
* Project Report – include the design of your system, e.g., term weighting and normalization schemes you use for query and documents, particular data structures/classes, and your system performance comparison among different query settings (e.g., precision and recall for each query under each setting). Different settings mean when you consider only the main query (title), when you consider the description along with the main query (description + title), and when you consider the narrative along with the main query (narrative + title).