CPSC5330 – Big Data Analytics
Lab 3: TFIDF (the end)

This is a continuation of the TFIDF search system we started in Lab3.

Lab3 focused on the Indexing (ingestion) phase, which took a set (corpus) of documents as input, and produced a set of parameters of the form `tfidf(doc_id, term)` for every document in the corpus, and every term in every document.

Now we move to building the query phase:

The Query Phase

The query phase is separate from the Index Phase. It is an interactive application, and might be launched as a console application or as a web service. It takes the TF-IDF calculations from indexing and uses that data to rank documents. The application accepts a line from the user with query words. The line is split into words and converted to terms just as in the index phase, so a query is a set of *terms.* Each document has a relevance score with respect to a query, defined as follows:

$$relevance(doc, Q) = \sum_{term \in Q} \frac{tfidf(doc, term)}{|Q|}$$

In response to a user query line, the application presents the user with the five most relevant documents, in descending order of relevance – or fewer if there are fewer than five documents with relevance > 0.

You should be careful to separate the system into three components:

- The UI – how to accept a query from the user and display the results
- The calculation – takes the user query, prepares the terms, and calcluates relevance based on TFIDF values from the ingestion phase. This module is called by the UI but is ignorant of how the UI is implemented; this module calls the storage module to get TFIDF values, but is ignorant of how TFIDF values are stored
- Storage: how are the TFIDF values stored. This module implements an interface called by the calculation. It will require random access to TFIDF value indexed by term and possibly by document.