Assignment Report 4020 A3

## Indexing Method

The indexing method that was used is a reverse indexer and it is a flat database with no hierarchies within the search engine. This mean during a query; the whole database is queried against rather than specific categories. I had used this method as the dataset is small and can easily be index and queried using this method. If a larger data set was presented, it would be beneficial to use Nested Document Indexing.

## Weighing Method

The method I used was TF-IDF as the scoring process. I used this to check how many times a term appeared within a document and gave it a score. This score and ranking was listed in the output file. TF-IDF was useful as I did not have any categories and importance in viewing how a score was weighted compared to the number of results.

## Program Design

The program has 3 distinct parts that work in unison with each other. Our first program converts the files from a compressed format to a decompressed format which allows to be read. Then there is a indexer which reads the files and converts it to a standard encoding. This encoding is "ISO-8859-1". This allows the indexer to then commit it to the index. The index is a single thread workload meaning that each entry needs to be entered before it moves onto the next. Finally, there is a searcher program; this program oversees scoring and outputting the result files into the same directory.