Programming on Lucene, Assignment 2

Kalyan and Strand

Introduction

The Aim of this lab was to implement a TF - IDF vector model, finally computing the cosine similarity between two documents.

Instruction Clarity

Instructions were clear. However, We felt that the steps to change the source code and compiling (Addding TermFreqVector) difficult to follow from the PDF. (We used ant to build our project, lost time since both of us were new to ant)

Implemenation

We succeeded at implementing every thing required for this lab. We also had changed the signatures for the function `normalize` to return values.

Difficulties

After writing the new code, we tested to see if it worked. We used IntelliJ Idea and Eclipse to run it, Eclipse gave us some errors that we couldn't figure out. We tried for quite some time to solve the problem and to find out what was wrong, but in the end we decided to try it in Intellij Idea / Android Studio. We also had difficulties initially setting the class path.

Experiments

We ran a couple with experiments with handcrafted dataset. We obtained this dataset from one of our theory exercises. With this we could esaily verify our results as we had few terms.

While manually conducting our experiments we observed that values of log10 were wrong. After after debugging we realized that inverse document frequencies had be casted to double or multipled with a float value to fix this issue. There was an issue with document frequency. Document frequency had an extra document. After debugging we realized that the lucene indexer had also accounted for hidden system files in the `index` folder, this was giving us wrong document frequencies. luke.

WIth different novels we notice different values for cosine similarity. For instance cosine simialrity between the document `27531-0` and `DarwinOriginofSpecies` is `0.061`.