Information retrieval, HW4

# Aharon Sharim, 052328523

# Ran Romano,

## General program structure

We've used Lucene's built in KNN classifier.

Given the train documents, we've added both the title and the body to the index, using the build in "English" analyzer to do stemming and stop-word removal.

After which we load the test documents as required and run each of them through an instance of Lucene's "KNearestNeighborDocumentClassifier", which uses the "more like this" similarity heuristic (trimming terms with low tf, or terms with low character count), to generate similarity between documents, using both the (default) BM25 similarity and the old, classif tf/idf similarity.

We've also attempted to experiment with the minDocFreq and minTermFreq parameter to the classifier, but from preliminary results on a partial sample of documents, didn't not find a meaningful improvement.

The program contains 5 main parts:

1. Configuration – reads and store the configuration file
2. MyDocument – a class to read & store the document data (id,title,body,category)
3. KnnClassifier – a class to create the classifier from a given MyDocument list and assign a class to a document
4. ResultAnalyzer – Given a list a document with real and assigned categories, calculate the Micro/Macro F1 scores
5. Main – the main body of the program to glue it all together

## Selecting optimal K

We've run the classification overnight using different K values from 1 up to 20, and calculated both the F1 micro score and F1 macro score for each K, using both the (default) BM25 similarity and the old, classif tf/idf similarity.

Attached here is a graph of the results:

As we can see, BM25 preforms better than the classic td/idf similarity across all tested K values.

Therefore, we've decided to use the default BM25 similarity in our code.

For BM25 similarity, we see a steady increase in accuracy up to K = 19, and then starts to level out, and increasing K does not increase accuracy anymore.

Therefore, in the case of BM25 similarity we conclude that the optimal K = 19.