**Information Retrieval And Web Search – Home Assignment 3**

**Hodaya Shalom <ID>, Michal Abramov 301834297**

**Abstract**

This report includes the decision we made in this assignment.

We chose to use **Lucene library functionality and Weka** library integrated with our implementations. <did we end up using Lucene?>

**Experiment Description**

We preformed our experiment as follows:

1. Reading the collection of documents and converting them to Weka format.

(note: files need to be written in Weka format only once, after they are written the first time this part can be skipped by specifying ‘recreateWekaDataFolders=false’ in ‘parameters.txt’)

1. Building classifier using the train set
2. Testing the classifier accuracy using the test set
3. Printing results to output file

This experiment was preformed using different k values to choose the value that gives us the best accuracy.

**Reading and converting documents to Weka format**

<what do we wants to say about this?>

**Building classifier**

<lemmatization?>

We used Weka’s filters, classifiers, Multifilter and FilteredClassifier.

In order to represent the documents as their tf-idf vectors we used Weka’s *StringToWordVector* filter, setting *TFTransform* and *IDFTransform*

<what did we use Recorder for?>

For the classifier we used *IBK* set to work with the KNN algorithm.

We used multifilter and FilteredClassifier to combine the *StringToWordVector*  filter and the *IBK* classifier.

<do we want to say anything about the options we used?>

After creating the classifier object, we build the classifier itself (classifier.buildClassifier) using the test data (training the classifier).

**Testing the classifier**

We test the classifier by:

1. for each instance to the test set:

* classifying instance using the classifier
* counting number of accurate classification and calculating the percentage of correct classifications

**Experiment results**

After running the experiment with different k value, the value that gave us the best result is <add selected value> that gave us accuracy percentage of <add accuracy>

<anything else we want to say about the results?>

<micro&macro F score?>

<lowercase?>