Find Duplicates Tool

This tool is designed to find duplicate text files in a folder. It does it using Lucene http://lucene.apache.org. By duplicate we mean files that are exactly the same or just slightly different.

How does Lucene detect duplicate files?

First Lucene creates and index of the words in each document that will be later used to evaluate the content of each of them. Once the index is created Lucene can compare each document to all the others in order to find the ones that are more similar to it. Lucene then selects the best candidate to be the representative of each group of similar documents in the folder by maximizing the scores of each document in the set. The document that maximizes the scores is chosen as the representative document for each set. Finally, Lucene evaluates the features of each document in each group and compares them between all the documents in each set.

The similarity between documents is computed according to how many of the document features match the features of the document chosen as representative for the set. In details we take the score assigned to each feature by Lucene, we sum only the features that match the ones in the candidate document for the set and then divide this value the result by the sum of the features in the candidate document.

The threshold for similarity is set experimentally at 80%. Documents that get a score over this value are considered duplicates of the candidate document.

Input

The tool requires in input the folder with the documents in plain text format to be evaluated, the folder where it can save the created index and the output file name.

Output

The output is a text file containing the candidate documents for each set of duplicates and the list of its duplicates. The features extracted by Lucene and used to compare the documents are also listed as well as the similarity percentage.