DMT-HOMEWORK1

REPORT FOR PART 2

1.) What values for 'r,' 'b' and 'n' did you choose?

We selected n=100 hash functions, mainly for running time purposes and also selected r=5, since we believe its adequate for quite big but not huge files, like the lyrics of a song. So, since b=n/r, we took 20 bands.

2.) The probability to have False-Negatives, in the set of candidate pairs, for the following Jaccard values: 0.88, 0.9, 0.95 and 1.

P(false negatives)= (1-J**r)**b

J=0.88	3.046294749697203e-07
J=0.9	1.7590987863524756e-08
J=0.95	1.2319241384161333e-13
J=1	0

P(false positives)= 1-(1-J**r)**b

J=0.85	0.9999919443933685
J=0.80	0.9996439421094793
J=0.75	0.995563693487102
J=0.70	0.9747805441880405
J=0.65	0.9151289183523188
J=0.60	0.8019024538382217

J=0.55	0.6439846948142496
J=0.50	0.4700507153168765

3.) How did you handle the presence of False-Positives in the set of candidate pairs to be Near-Duplicates?

It is sure that we will have false positives in our candidates results, based on the values of r and b that we chose, however, we decided to emphasize on limiting the false negatives for this homework. We could try to increase r and minimize b in order to have smaller false positives, but then we would have bigger false negatives, due to the tradeoff between them. For this specific homework, our conception was to minimize the false-negatives.

4.) How did you reduce the probability to have False-Negatives?

By selecting a relatively small r, we are very confident that we will have no false negatives for jaccard similarity 0.88 or more.

5.) The Execution-Time of the Near-Duplicates-Detection tool.

2 minutes and 12 seconds

6.) The number of Near-Duplicates couples you found.

27.120