Problem 1: Evaluation (20pt)

A. The following list of R’s and N’s represents relevant (R) and non-relevant (N) documents in a ranked list of 50 documents. The top of the ranked list is on the left of the list. They are the documents that the system believes are most likely to be relevant. This list shows 8 relevant documents that are retrieved. Assume that there are an additional 2 relevant documents that were not retrieved by this system. Based on the list

- Compute the average precision.
- Compute and plot the interpolated precision for the standard 11 recall points

RRRRNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN

B. Now, consider another system that retrieves the following ranked list for the same query (thus the number of relevant documents is 10). Repeat the problem A by (a) computing the average precision, and (b) computing the interpolated precision for 11 recall points. What do the curves tell you about the system in prob. A and the one in prob. B?

RRRRNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN

Problem 2 (10pt): Stopword List

Examine the Lemur stopword list (http://www.cse.msu.edu/~cse484/hw/stopwords) and list 10 words that you think would cause problems for some queries. Give these 10 words and explain why they cause problems.

Problem 3 (10pt): Porter Stemming

Download the C version of the Porter stemming algorithm from http://tartarus.org/~martin/PorterStemmer/c.txt and the java version from http://tartarus.org/~martin/PorterStemmer/java.txt. Run the Porter stemming algorithm against the following words: ironic, iron, animal, animated, animation,
policy, police, organ, organization, organize, consist, consisted, consistency, consistent, connection, connections, connected, connecting, computer, compute, computers, computation, argue, argument, arguing, general, generally, generation, generations, generative, generous, generals, sever, several, severe, severing, severity, severs. Identify the group of words that share the same word stems generated by the Porter stemming algorithm. You need to submit in this homework (i) the word stems generated by Porter stemming algorithm and the associated groups of words, (ii) the discussion of which word group makes sense and which does not.