Information Retrieval Evaluation – Exercise 2
IST 441

This exercise is worth 6 points.

In this exercise, you will familiarize yourself with standard methods for evaluating information retrieval systems with an emphasis on precision, recall and F1.

1. (4 points) Suppose that an IR system contains only 1000 documents numbered d_1 to d_{1000}. A query is known to generate 27 relevant documents as listed below:

{d_1, d_5, d_7, d_{10}, d_{88}, d_{151}, d_{200}, d_{211}, d_{250}, d_{300}, d_{399}, d_{401}, d_{405}, d_{450}, d_{473}, d_{500}, d_{501}, d_{530}, d_{545}, d_{590}, d_{600}, d_{735}, d_{700}, d_{720}, d_{800}, d_{888}, d_{900}}.

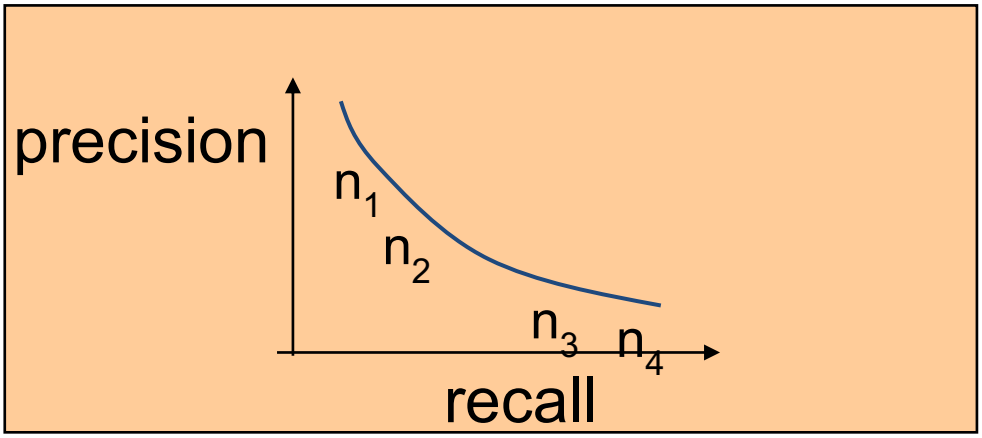
Two very different IR systems are used to retrieve ranked documents for this query. Each system only returns the top 10 ranked documents in the following order of ranking. Systems 1 and 2 each retrieves documents one at a time in the following order with all 10 documents eventually returned:

System 1: d_{122}, d_{211}, d_{150}, d_{88}, d_{37}, d_1, d_{501}, d_{800}, d_{201}, d_5.

System 2: d_{10}, d_6, d_{700}, d_{250}, d_{88},, d_{600}, d_{59}, d_{422}, d_{500}, d_{333}.

Answer the following and show your work:

- Plot the Precision and the Recall graphs for each system as a function of the number of documents returned (for 1 document returned, 2 documents returned, etc).
- Plot the Precision versus Recall for systems 1 and 2 using these query results as a function of the number of documents returned. Note that n_1 is the value of precision and recall for the first document, n_2 for the 2 documents.
- Calculate the F1 score for each.
- Which IR system is better? Justify your answer.



2. (2 points) What can be measured by a search engine? Specify the different kinds of search engines that one could have or use. Can one measure precision, recall, and F1? Why?