Cerca i Anàlisi d'Informació Massiva (CAIM)
Grau en Enginyeria Informàtica (GEI)

Session 4: Evaluation

Exercise List, Fall 2018

Basic comprehension questions.
Check that you can answer them before proceeding. Not for credit.

1. Explain why reference collections are useful in Information Retrieval, and how they are used.
2. Write down the definitions of recall, precision, coverage, and novelty. Explain them in words in a way that you think your classmates would understand.
3. Explain to yourself how to compute a precision/recall graph.
4. True or false or criticize: To maximize user satisfaction, aim at a balance between recall and precision
5. Define a snippet to yourself.

Note: A spreadsheet may help a lot in solving some of the exercises below.

Exercise 1

A user tells us that, after asking a query to our search system, she found 10 relevant documents in positions 2, 6, 12, 18, 20, 22, 30, 36, 40, and 50. Assuming there are no more relevant documents in the collection, draw a precision-recall graph of the answer at 10 recall levels. Make sure you give the table of numbers that you used to plot the graph.

Exercise 2

We have a document collection with 100 documents, identified by numbers $1 \ldots 100$. Suppose that the relevant ones for a given query are those numbered $1 \ldots 20$.

Two information retrieval systems give as a result to the query the following answers:

S1= { 1,2,21,22,3,23,25,4,28,5,29,30,6,7,31,32,33,40,41,42,8,43,44,
9,45,10,50,51,11,52,53,54,12,60,62,13,63,64,14,15,16,70,78,80,17,
81,82,83,85,18,90,19,91,92,20,93,94,95,96,98 },

S2= { 25,26,1,27,28,2,3,29,30,4,35,36,5,37,6,7,8,38,9,40,10,42,11,45,46,
12,48,50,51,13,60,61,64,14,70,72,15,78,79,90 }.

For this query and each of the two systems,

a) Compute the recall, precision, and F-measure (with $\alpha = 1/2$, $\alpha = 1/4$, and $\alpha = 3/4$).

b) Compute the novelty and coverage measures, assuming that the user already knew the documents with odd index and did not knew about those with even index.

Exercise 3

For the collection, query, and system S1 in the previous exercise,

a) Give plots showing % recall, precision, and interpolated precision as a function of the number of retrieved documents.

b) Give the 11-point precision-recall graph.

c) Compute the average precision at those 11 points.

d) Compute (approximately) the AUC (Area under Curve), with respect to recall. To do this, compute the area of 10 rectangles of base 10% each in the recall-rank plot, then normalize dividing by the maximum possible area, i.e., the one that a perfect query would cover.

Exercise 4

Suppose that, in a document collection containing D documents, there are R of them that are really relevant for a query issued by a user. Describe as precisely as possible:

- the precision/recall graph of a perfect query, one that returns the R relevant documents in the top-R positions of the ranking and may contain non-relevant documents starting at position $R + 1$;

- the precision/recall graph of a totally uninformed query, one that returns as answer a list of R randomly selected documents from the whole; what is the area under the curve of this plot?

Exercise 5

We introduce a new measure of query effectiveness called *correctness*. It is defined as the fraction of documents that are correctly labeled by the system, i.e., relevant and in the answer or else irrelevant and not in the answer.

Suppose we have d documents. For a given query, suppose that there are $r \leq d$ relevant documents, and our system answers with a relevant documents and b irrelevant documents ($a + b \leq d$). For example, if $d = 1000$, $r = 100$, $a = 60$, and $b = 20$, we have recall = 60%, precision = 75% and correctness = 94%.

1. Define correctness in terms of d, r, a and b. Define it in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

2. Fix a query q and $d = 1000$. Give triples of values r, a, and b such that

   - correctness and recall are high and precision is low
   - corectness and precision are high and recall is low
   - correctness and precision are low and recall is high

   (as a reference, think of "high" as $> 80\%$ and "low" as $< 20\%$).

3. Argue that it is not possible to have low correctness but both high recall and high precision. To do this: Express "low correctness" as a condition on TN, TP, FN, FP. Deduce that either FP or FN is relatively high. Conclude that then either recall is not high or precision is not high. (Note: Do not just repeat this informal argument. Make it rigorous.)