Basic comprehension questions.
Check that you can answer them without looking at the course slides. If you can’t, read the slides and then try again.

1. Hide the course slides. Write down the formula for recall and precision. Also, explain them in words in a way that you think your classmates would understand.

2. Same as above for Rocchio’s relevance feedback formula.

3. Explain to yourself how to compute a precision-recall plot.

4. Tell the difference between user relevance feedback and pseudorelevance feedback.

5. True or false or criticize: To maximize user satisfaction, aim at a balance between recall and precision.

Exercise 1
We have a document collection with 100 documents, identified by numbers 1...100. Suppose that the relevant ones for a given query are those numbered 1...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, compute recall and precision.

**Exercise 2**

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 3**

For the collection, query, and system S1 in exercise 1:

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 the AUC of this query with respect to recall. The AUC, or Area under the Curve, is a measure of how well a query does compared to what an “optimal” query would do. To do this, first imagine what an optimal query would look like: one that first gets all relevant document and maybe after that starts giving irrelevant documents. Compute the area that would be below that its curve. Then, compute (approximately) the area under the curve for this particular query: the area of 10 rectangles of base 10% each in the recall-rank plot. Then, divide this number by the area under the perfect query. The result is a number between 0 and 1. The closer to 1, the better the query.

**Exercise 4**

We introduce a new measure of query effectiveness called correctness. It is defined as the fraction of documents that are well-labelled 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. Say why you think correctness is not very interesting in Information Retrieval systems. (In contrast, in fields such as machine learning, correctness is the main measure when evaluating for example a classifier algorithm (with 2-class problems.).)

3. 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
   • correctness 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 $< 20\%$).

4. 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.)

**Exercise 5**

We have indexed a document collection, and included the following terms in the vocabulary. The second column indicates the percentage of documents where each term appears.

<table>
<thead>
<tr>
<th>Term</th>
<th>% docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer</td>
<td>10%</td>
</tr>
<tr>
<td>software</td>
<td>10%</td>
</tr>
<tr>
<td>bugs</td>
<td>5%</td>
</tr>
<tr>
<td>code</td>
<td>2%</td>
</tr>
<tr>
<td>developer</td>
<td>2%</td>
</tr>
<tr>
<td>programmers</td>
<td>2%</td>
</tr>
</tbody>
</table>

The documents are represented internally as vectors using tf-idf weights.

Given the query $Q$=“computer software programmers”, an information retrieval system returns documents $D1$=“a computer is useless without software” $D2$=“programmers spent much of their time finding bugs in code”, and
D3 = “programmers are usually good thinkers”. The user considers all of $D_1$, $D_2$, and $D_3$ relevant.

Suppose the system implements user relevance feedback using Rocchio’s rule with $\alpha = 0.8$, $\beta = 0.2$, and $\gamma = 0$. Give, as a vector, the new query built by the system at this point.