Evaluation in Information Retrieval (Chapter 8)

Definition 1 (Recall)
Recall describes how many of the relevant documents are retrieved.

$$recall = R = \frac{\#relevant\ retrieved}{\#relevant}$$

Definition 2 (Precision)
Precision describes how many of the retrieved documents are relevant.

$$precision = P = \frac{\#relevant\ retrieved}{\#retrieved}$$

Definition 3 (*F*-measure)
A balanced F-measure (F_1-measure) defines a recall-precision relationship represented by their weighted harmonic mean:

$$F = \frac{2 \cdot R \cdot P}{R + P}$$

Definition 4 (Mean Average Precision)
MAP expresses the precision in each point a new relevant document is included in the result. It is counted as

$$MAP(Q) = \frac{1}{|Q|} \cdot \left(\sum_{q \in Q} \frac{1}{rel_q} \cdot \left(\sum_{i=1}^{rel_q} prec_i \right) \right)$$

where rel_q is the number of relevant documents for query q and $prec_i$ is the precision at the i-th relevant document.

Definition 5 (κ statistic)
Let N be the total number of documents, J is a set of judges and $P(A) = \frac{\#agree}{N}$ the number of documents on which the judges agree. Let also define R_j and NR_j be the number of relevant and non-relevant documents, respectively, according to the judge $j \in J$ and

$$P(R) = \frac{\sum_{j \in J} R_j}{|J| \cdot N} \quad \text{and} \quad P(NR) = \frac{\sum_{j \in J} NR_j}{|J| \cdot N}$$

as the number of relevant and non-relevant documents, respectively. Let finally define

$$P(E) = P(R)^2 + P(NR)^2$$

as the approximate number of disagreements between the judges. Then the κ statistic is defined as the measure of agreement between the judges

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}.$$

Exercise 8/1

The following ordered list of 20 letters R and N represents relevant (R) and non-relevant (N) retrieved documents as an answer for a query on a collection of 10 000 documents. The leftmost document is expected to be the most relevant. The list contains 6 relevant documents. Assume that the collection contains 8 documents relevant to the query.

$$RRNNNNNNRNRNNNRNNNNR$$

a) What is the precision on the first 20 results?

b) What is the F-measure on the first 20 results?

c) What is the non-interpolated precision of the system at 25% recall? (R=25%)

d) What is the interpolated precision of the system at 33% recall? (R>33%)

e) Assume that these 20 documents are the complete list of retrieved documents. What is the MAP of the system?

Now assume that the system returned all 10,000 documents in an ordered list and above is the top 20.

f) What is the highest possible MAP the system can achieve?

g) What is the lowest possible MAP the system can achieve?

Applying the Definition 2 do calculate (a) we get $P = \frac{6}{20} = \frac{3}{10}$. For (b) it is necessary to count the recall with Definition 1 as $R = \frac{6}{8} = \frac{3}{4}$. Using Definition 3 with $\alpha = 0.5$ we count

$$\frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{(1^2+1)\cdot\frac{3}{10}\cdot\frac{3}{4}}{\frac{3}{10}+\frac{3}{4}} = \frac{\frac{9}{20}}{\frac{21}{20}} = \frac{3}{7}.$$

To count the non-interpolated precision of the system at 25% recall for (c) we need the precisions for the documents of recall equal to 25%:

$$\begin{aligned}
&1. \quad P = \tfrac{1}{1} \qquad R = \tfrac{1}{8} \quad = 12.5\% \\
&2. \quad P = \tfrac{2}{2} \qquad R = \tfrac{2}{8} \quad = 25\% \\
&3. \quad P = \tfrac{2}{3} \qquad R = \tfrac{2}{8} \quad = 25\% \\
&\qquad\quad \cdots \\
&8. \quad P = \tfrac{2}{8} \qquad R = \tfrac{2}{8} \quad = 25\% \\
&9. \quad P = \tfrac{3}{9} \qquad R = \tfrac{3}{8} \quad = 37.5\%
\end{aligned}$$

We see that the first document has $R = 12.5\%$ but this value is less than 25%. Documents 2 to 8 have the desired recall of 25%. Document 9 has already a higher value so we do not include it in the result. Non-interpolated precision is then the set of precisions of these 7 documents

$$P = \left\{ \frac{2}{2}, \frac{2}{3}, \frac{2}{4}, \frac{2}{5}, \frac{2}{6}, \frac{2}{7}, \frac{2}{8} \right\}.$$

For the interpolated precision (d) we are looking for the highest precision for the relevant documents of recall higher than 33%. Recall changes if and only if the result contains a relevant document. Therefore, the values are calculated for the documents 11, 15 and 20.

$$\begin{aligned}
&11. \quad P = \tfrac{4}{11} \qquad R = \tfrac{4}{8} \quad = 50\% \\
&15. \quad P = \tfrac{5}{15} \qquad R = \tfrac{5}{8} \quad = 62.5\% \\
&20. \quad P = \tfrac{6}{20} \qquad R = \tfrac{6}{8} \quad = 75\%
\end{aligned}$$

The requested recall value is exceeded by retrieving the document 9 (37.5 %). Now we only have to find $max\{P_9, P_{11}, P_{15}, P_{20}\} = max\{\frac{3}{9}, \frac{4}{11}, \frac{5}{15}, \frac{6}{20}\} = \frac{4}{11} = 0.\overline{36}$.

To estimate the MAP of the system (e) we use the Definition 4. Since we only have one query $N = |Q| = 1$ and the first 20 documents contain $rel_q = 8$ relevant documents

$$MAP(Q) = \frac{1}{1}\left(\sum_{j=1}^{1}\frac{1}{8}\left(\sum_{i=1}^{6}P(doc_i)\right)\right) = \frac{1}{1}\cdot\frac{1}{8}\left(\underbrace{\frac{1}{1}}_{P(1.)} + \underbrace{\frac{2}{2}}_{P(2.)} + \underbrace{\frac{3}{9}}_{P(9.)} + \underbrace{\frac{4}{11}}_{P(11.)} + \underbrace{\frac{5}{15}}_{P(15.)} + \underbrace{\frac{6}{20}}_{P(20.)}\right)$$

$$= \frac{1}{1}\cdot\frac{1}{8}\cdot\frac{1099}{330} = \frac{1099}{2640} = 0.4162\overline{87}$$

From Definition 2 and Definition4 follows that if the two remaining relevant documents were on positions 21 and 22 then MAP with $rel_q = 8$ relevant documents is the highest possible (f)

$$MAP(Q) = \frac{1}{8}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{21} + \frac{8}{22}\right) = 0.503\overline{409}$$

and, on the other hand, if they were on the last places the MAP would be minimal (g)

$$MAP(Q) = \frac{1}{8}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{9999} + \frac{8}{10000}\right) = 0.4164\overline{7538}.$$

Exercise 8/2

The following ordered list of 5 letters R and N represent relevant (R) and non-relevant (N) retrieved documents as an answer for a query on a collection of 100 documents. The leftmost document is expected to be the most relevant. The list contains 3 relevant documents. Assume that the collection contains 5 documents relevant to the query.

$$RNNRR$$

a) What is the F-measure on the first 5 results?

Assume that these 5 documents is the complete list of retrieved documents.

b) What is the MAP of the system?

Now assume that the system returned all 100 documents in an ordered list and above is the top 5.

c) What are the highest and lowest possible MAPs the system can achieve?

---

a) $R = \frac{3}{5}$, $P = \frac{3}{5}$, then $F = \frac{3}{5}$.

b) $\frac{1}{5}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5}\right)$.

c) highest $\frac{1}{5}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5} + \frac{4}{6} + \frac{5}{7}\right)$ and lowest $\frac{1}{5}\left(\frac{1}{1} + \frac{2}{4} + \frac{3}{5} + \frac{4}{99} + \frac{5}{100}\right)$.

Exercise 8/3

The following two sequences of letters R and N represent the complete lists of relevant (R) and non-relevant (N) retrieved documents as answers for two queries on a collection of 100 documents. The leftmost document is expected to be the most relevant. Assume that the collection contains 10 documents relevant to the first query and 20 documents relevant to the second query. Find the F-measure and the MAP of this system.

$$NRNNNRN \quad \text{and} \quad NNRRR$$

- $R_1 = \frac{2}{10}$, $P_1 = \frac{2}{7}$, then $F_1 = \frac{\frac{4}{35}}{\frac{17}{35}} = \frac{4}{17}$.
- $R_2 = \frac{3}{20}$, $P_2 = \frac{3}{5}$, then $F_2 = \frac{\frac{9}{50}}{\frac{3}{4}} = \frac{6}{25}$.
- $F = \frac{F_1+F_2}{2} = \frac{101}{425}$.
- $\text{MAP}_1 = \frac{1}{10}\left(\frac{1}{2} + \frac{2}{6}\right) = \frac{3+2}{2*6} = \frac{1}{12}$.
- $\text{MAP}_2 = \frac{1}{20}\left(\frac{1}{3} + \frac{2}{4} + \frac{3}{5}\right) = \frac{20+30+36}{20*60} = \frac{43}{600}$.
- $\text{MAP} = \frac{\text{MAP}_1+\text{MAP}_2}{2} = \frac{93}{1200}$.

Exercise 8/4

Below is a table showing how two judges judged the relevance (0 = non-relevant, 1 = relevant) of the set of 12 documents with respect to a query. Assume that you developed an IR system, that for this query returns the documents $\{4, 5, 6, 7, 8\}$.

Doc ID	Judge 1	Judge 2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

Table 1: Judges judging the relevance of documents.

a) Calculate the κ statistic.

b) Calculate the recall, precision and F-measure of your system in which a document is considered relevant if the judges agree.

c) Calculate the recall, precision and F-measure of your system in which a document is considered relevant if at least one of the judges thinks so.

For (a) it is necessary to use the Definition 5. First we count $P(A)$ as the number of documents on which the judges agree. Since these are the documents $\{1,2,3,4\}$ and the total number of them is $N = 12$, then $P(A) = \frac{|\{1,2,3,4\}|}{12} = \frac{4}{12} = \frac{1}{3}$. Now we need the counts of disagreements between the judges. Judge 1 considers the documents $NR_1 = \{1,2,9,10,11,12\}$ and judge 2 $NR_2 = \{1,2,5,6,7,8\}$ as non-relevant. Plugging in to the formula for $P(NR)$ we get $P(NR) = \frac{|NR_1|+|NR_2|}{2\cdot 12} = \frac{2\cdot 6}{24} = \frac{1}{2}$. We repeat this for $P(R)$. Since the number of relevant is equal to non-relevant, then $P(R) = P(NR) = \frac{1}{2}$. We finally can count $P(E)$ as

$$P(E) = P(NR)^2 + P(R)^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

and

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{\frac{1}{3} - \frac{1}{2}}{1 - \frac{1}{2}} = -\frac{1}{3}.$$

If $\kappa < 0$ then the disagreement between the judges is more than random.

For (b) it is necessary to calculate recall and precision. Our system retrieves the documents $\{4,5,6,7,8\}$ as relevant whereas the judges only agree on $\{3,4\}$. The intersection is $\{4\}$. As to the Definition 2 we obtain

$$P = \frac{|\{4\}|}{|\{4,5,6,7,8\}|} = \frac{1}{5}$$

and, as the number of relevant documents is $|\{3,4\}| = 2$, the recall is $R = \frac{1}{2}$ by Definition 1. Then, by Definition 3 the F-measure is equal to

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{(1+1)\frac{1}{5}\frac{1}{2}}{\frac{1}{5} + \frac{1}{2}} = \frac{2}{7}.$$

We similarly count these for (c) which says that a document is relevant if at least one judge considers it relevant. This makes the documents $\{3,4,5,6,7,8,9,10,11,12\}$ relevant. Their intersection with our result $\{4,5,6,7,8\}$ is the set $\{4,5,6,7,8\}$ of size 5. Recall is $R = \frac{|\{4,5,6,7,8\}|}{|\{3,4,5,6,7,8,9,10,11,12\}|} = \frac{5}{10} = \frac{1}{2}$, precision is $P = \frac{5}{5} = 1$ and F-measure is

$$F = \frac{(1+1)\frac{1}{2}}{1 + \frac{1}{2}} = \frac{2}{3}.$$