Exercise 8.1, 8.2

- An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?
- The balanced F measure (a.k.a. F1) is defined as the harmonic mean of precision and recall. What is the advantage of using the harmonic mean rather than "averaging" (using the arithmetic mean)?

Exercise 8.5, 8.6, 8.7

- 8.5: Must there always be a break-even point (p=r) between precision and recall?
- 8.6: What is the relationship between the value of *F*1 and P and R at the break-even point?
- 8.7: The *Dice coefficient* of two sets is a measure of their intersection scaled by their size (giving a value in the range 0 to 1):

$$\text{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

  - Show that the balanced F-measure (*F*1) is equal to the Dice coefficient of the retrieved and relevant document sets.

Exercise 8.8

- Consider an information need for which there are 4 relevant documents in the collection.
- Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):
  - System 1: R N R N N N N N R R
  - System 2: N R N N R R R N N N
- a. What is the MAP of each system? Which has a higher MAP?
- b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
- c. What is the R-precision of each system? (Does it rank the systems the same as MAP?)

Exercise 8.9

- The following list of R's and N's represents relevant (R) and nonrelelvant (N) returned documents in a ranked list of 20 docs retrieved by a query from a collection of 20,000 documents. Assume that there are 8 relevant documents in total in the collection.
  - RRNNN NNNRN RNNNR NNNNR
- a) What is the Precision on top 20?
- b) What is the F1 on top 20?
- c) What is the unintepolated precision at 25% recall?
- d) Plot the interpolated precision up to recall 75%
- e) What is the largest possible MAP that the system may have? (if the remaining 19,980 docs are ranked after these top 20)
- f) What is the smallest possible MAP that the system may have?

Exercise 8.10

- The table shows how two judges rated the relevance of 12 documents (0 = nonrelevant, 1 = relevant) for a query.
- Let us assume that you've written an IR system that for this query returns the set of documents {4, 5, 6, 7, 8}.
- a. Calculate the kappa measure between the two judges.
- b. Calculate precision, recall, and *F*1 of your system if a document is considered relevant only if the two judges agree.
- c. Calculate precision, recall, and *F*1 of your system if a document is considered relevant if either judge thinks it is relevant.

docID	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