Information Retrieval

INFO 4300 / CS 4300

- Evaluation
  - Evaluation corpus and logging
  - → Metrics
  - Training, testing

Ranking Effectiveness

= the relevant documents

Ranking #1

| Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

Precision at rank p: for p=10, prec@10=0.6 for both rankings.

Summarizing a Ranking

- Options
  - Calculating **recall** and **precision at fixed rank positions** (which need to be specified)
  - Calculating **precision at standard recall levels**, from 0.0 to 1.0
    - » requires *interpolation*
  - *Average* the *precision* values **from the rank positions where a relevant document was retrieved**

Average Precision

= the relevant documents

Ranking #1

Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

Ranking #1: $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$

Ranking #2: $(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$

Averaging Across Queries

▮▮▮▮▮ = relevant documents for query 1

Ranking #1 ▮□▮□□▮□□▮▮

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

▮▮▮ = relevant documents for query 2

Ranking #2 □▮□□▮□▮□□□

Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3
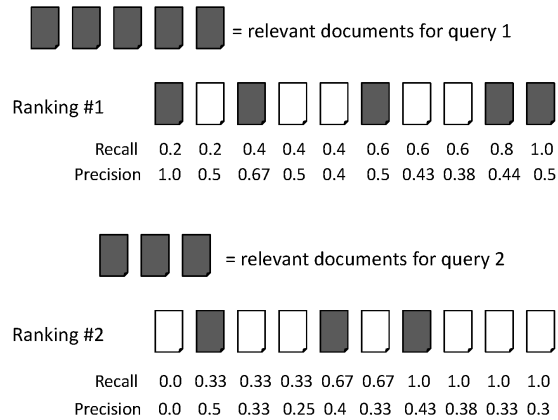
Averaging

- ***Mean Average Precision*** (MAP)
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers
  - assumes user is interested in finding many relevant documents for each query
  - requires many relevance judgments in text collection
- **Recall-precision graphs** are also useful summaries

MAP

▮▮▮▮▮ = relevant documents for query 1

Ranking #1 ▮□▮□□▮□□▮▮

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

▮▮▮ = relevant documents for query 2

Ranking #2 □▮□□▮□▮□□□

Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3
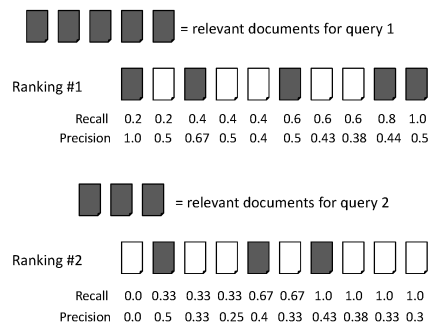
$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

Recall-Precision Graph

- For the 2 queries from the previous slide
- All of the R/P values available should be included in the graph
- Gives more detail on the effectiveness at different recall levels
- To generate a single graph that summarizes all queries requires interpolation
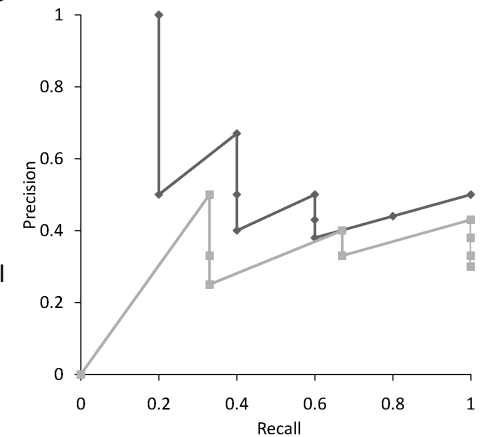
To average the graphs...

- Calculate precision at **standard recall levels**
  - 0.0 to 1.0 in increments of 0.1
- Requires interpolation
  - Define a function to map the existing R/P points onto R/P points for the standard recall levels
- In IR, we rely on the fact that precision tends to decrease as recall increases
  - Produce a function that is monotonically decreasing
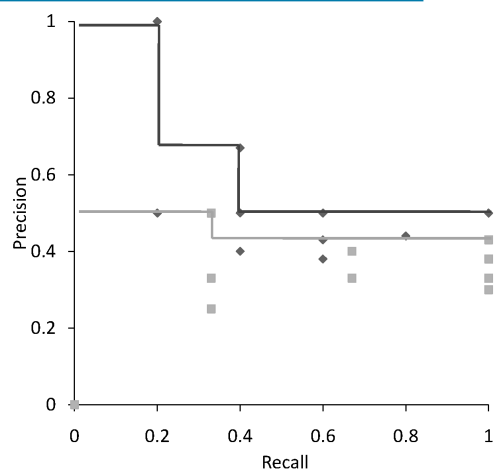
Interpolation

- Calculate precision at each standard recall level R as the **maximum** precision observed in any recall-precision point at a higher recall level:

$$P(R) = \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

  - where S is the set of observed (R, P) points
- Produces a step function
- Defines precision at recall 0.0

Interpolation

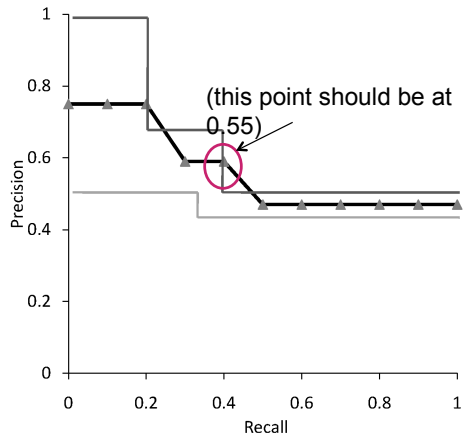- Interpolated R/P graph for two queries in the MAP example

Average Precision at Standard Recall Levels

- At each standard recall level, take the average of the precision values for each query

Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Ranking 1	1.0	1.0	1.0	0.67	0.67	0.5	0.5	0.5	0.5	0.5	0.5
Ranking 2	0.5	0.5	0.5	0.5	0.43	0.43	0.43	0.43	0.43	0.43	0.43
Average	0.75	0.75	0.75	0.59	0.55	0.47	0.47	0.47	0.47	0.47	0.47

Average Recall-Precision Graph

- Recall-precision graph plotted by simply joining the average precision points at the standard recall levels

(this point should be at 0.55)

Graph for 50 Queries