NLP and IR

Lab02: QA Retrieval & Evaluation

Aliaksei Severyn

University of Trento, Italy

March 28, 2013

Plan for the lab

- QA retrieval
- Evaluation metrics
  - MRR, Recall@N, Accuracy, MAP
- Preprocessing
  - Standard Analyzer vs. Stemming
  -

Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

F1

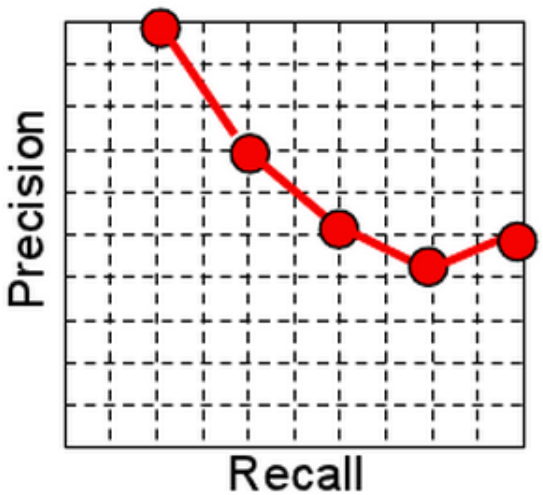
The weighted harmonic mean of precision and recall

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

Metrics for ranked results

- **Precision** and **recall** are well-defined for sets
- For ranked retrieval:
  - compute recall and precision

  at each rank
  - plot precision vs. recall
- MRR
- Average Precision:
  - **average precision** at ranks where relevant documents occurred



Mean Reciprocal Rank

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}.$$

Query	Results	Correct response	Rank	Reciprocal rank
cat	catten, cati, **cats**	cats	3	1/3
torus	torii, **tori**, toruses	tori	2	1/2
virus	**viruses**, virii, viri	viruses	1	1

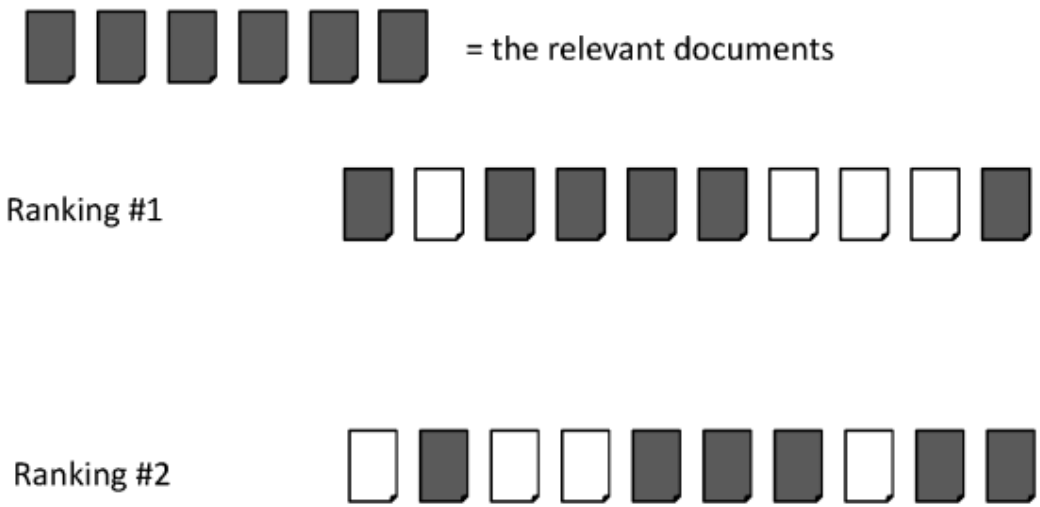
(1/3 + 1/2 + 1)/3 = 11/18 = 0.61

Mean Average Precision

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \mathrm{rel}(k))}{\text{number of relevant documents}}$$

Example: Average Precision



Ranking #1:

Ranking #2:

Example: 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

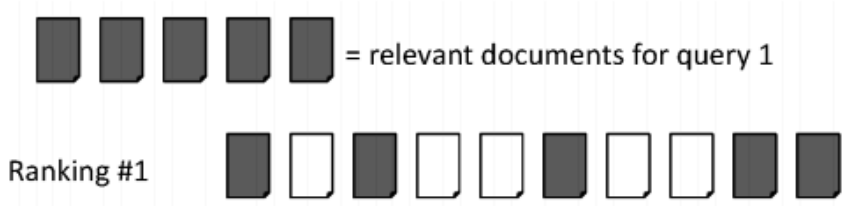
Ranking #2

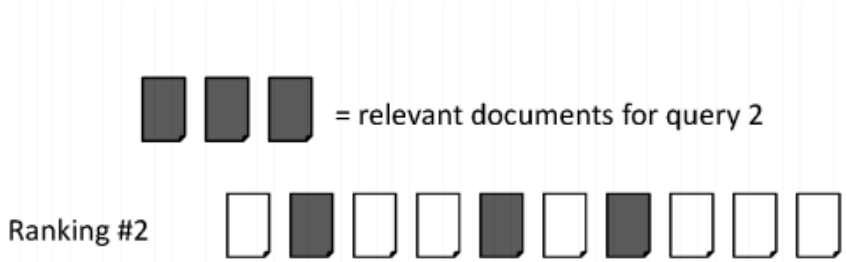
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

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

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

Example: MAP

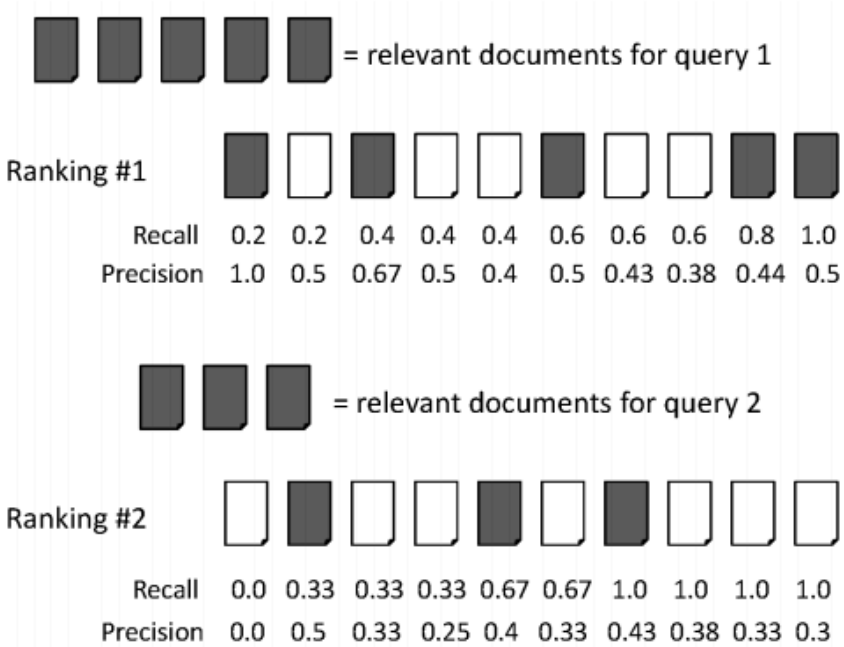




$$average\ precision\ query\ 1 =$$
$$average\ precision\ query\ 2 =$$

$$mean\ average\ precision =$$

Example: MAP



$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

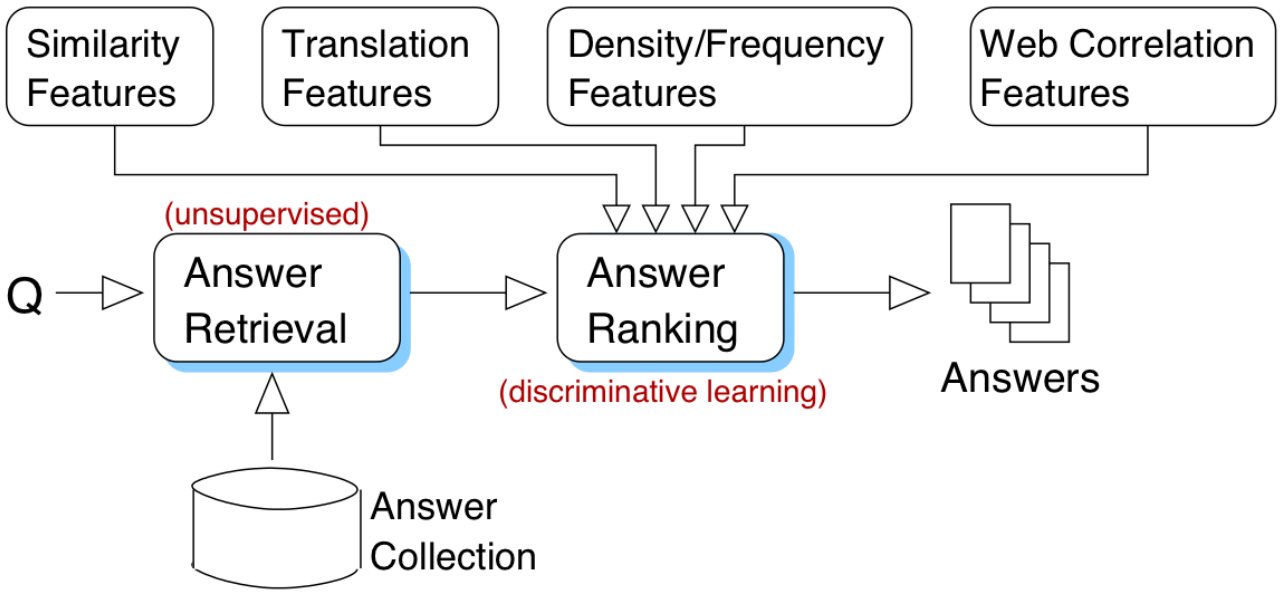
$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

MAP summary

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero
- MAP is macro-averaging: each query counts equally
- One of the most commonly used measures in research papers
- MAP assumes user is interested in finding many relevant documents for each query

Evaluating our QA

Simple QA pipeline



Using Standard Analyzer

```
        IR
MRR: 66.52
MAP:  0.67
REC-1@01:  57.39  ACC@01:  57.39  AC1@01:   0.67  AC2@01: 2869
REC-1@02:  68.03  ACC@02:  34.02  AC1@02:   0.79  AC2@02: 3401
REC-1@03:  73.25  ACC@03:  24.42  AC1@03:   0.85  AC2@03: 3662
REC-1@04:  76.06  ACC@04:  19.01  AC1@04:   0.88  AC2@04: 3802
REC-1@05:  78.32  ACC@05:  15.66  AC1@05:   0.91  AC2@05: 3915
REC-1@06:  79.74  ACC@06:  13.29  AC1@06:   0.92  AC2@06: 3986
REC-1@07:  80.76  ACC@07:  11.54  AC1@07:   0.94  AC2@07: 4037
REC-1@08:  81.96  ACC@08:  10.24  AC1@08:   0.95  AC2@08: 4097
REC-1@09:  82.74  ACC@09:   9.19  AC1@09:   0.96  AC2@09: 4136
REC-1@10:  83.64  ACC@10:   8.36  AC1@10:   0.97  AC2@10: 4181
REC-1@11:  84.38  ACC@11:   7.67  AC1@11:   0.98  AC2@11: 4218
REC-1@12:  84.90  ACC@12:   7.07  AC1@12:   0.98  AC2@12: 4244
REC-1@13:  85.38  ACC@13:   6.57  AC1@13:   0.99  AC2@13: 4268
REC-1@14:  85.84  ACC@14:   6.13  AC1@14:   1.00  AC2@14: 4291
REC-1@15:  86.22  ACC@15:   5.75  AC1@15:   1.00  AC2@15: 4310
```

With Stemming

```
        IR
MRR: 68.74
MAP:  0.69
REC-1@01:  59.03  ACC@01:  59.03  AC1@01:   0.66  AC2@01: 2951
REC-1@02:  70.49  ACC@02:  35.25  AC1@02:   0.79  AC2@02: 3524
REC-1@03:  76.18  ACC@03:  25.39  AC1@03:   0.85  AC2@03: 3808
REC-1@04:  79.02  ACC@04:  19.75  AC1@04:   0.89  AC2@04: 3950
REC-1@05:  81.28  ACC@05:  16.26  AC1@05:   0.91  AC2@05: 4063
REC-1@06:  82.78  ACC@06:  13.80  AC1@06:   0.93  AC2@06: 4138
REC-1@07:  83.96  ACC@07:  11.99  AC1@07:   0.94  AC2@07: 4197
REC-1@08:  84.96  ACC@08:  10.62  AC1@08:   0.95  AC2@08: 4247
REC-1@09:  85.90  ACC@09:   9.54  AC1@09:   0.96  AC2@09: 4294
REC-1@10:  86.84  ACC@10:   8.68  AC1@10:   0.97  AC2@10: 4341
REC-1@11:  87.26  ACC@11:   7.93  AC1@11:   0.98  AC2@11: 4362
REC-1@12:  87.76  ACC@12:   7.31  AC1@12:   0.98  AC2@12: 4387
REC-1@13:  88.24  ACC@13:   6.79  AC1@13:   0.99  AC2@13: 4411
REC-1@14:  88.66  ACC@14:   6.33  AC1@14:   0.99  AC2@14: 4432
REC-1@15:  89.16  ACC@15:   5.94  AC1@15:   1.00  AC2@15: 4457
```

Next time

- SVM Reranker
  - BOW
  - Shallow structures