Week 3

Querying and ranking: Measuring the quality

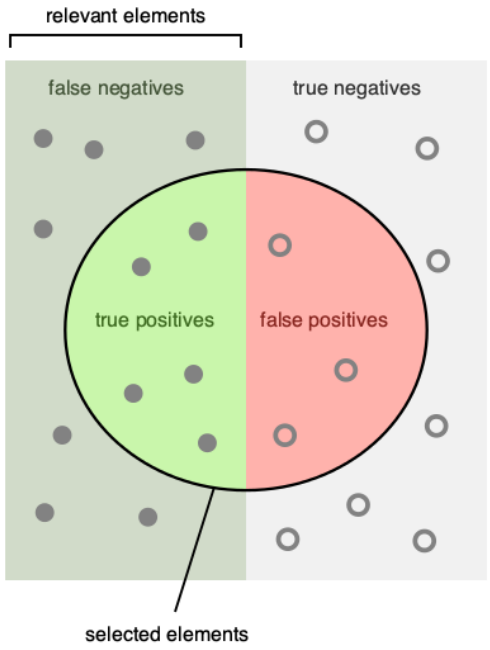
Goran Nenadic

with examples from the IIR book

Measuring quality of retrieval

- How do we measure if the retrieval was successful?
- "Outcomes" - contingency table

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)



tp = retrieved and relevant
fp = retrieved but not relevant
fn = not retrieved but relevant
tn = not retrieved and not relevant

Increase *tp* and *tn*, decrease *fp* and *fn*

https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Measuring quality of retrieval

- Several measures including:
  - **Precision**: fraction of retrieved documents that are relevant

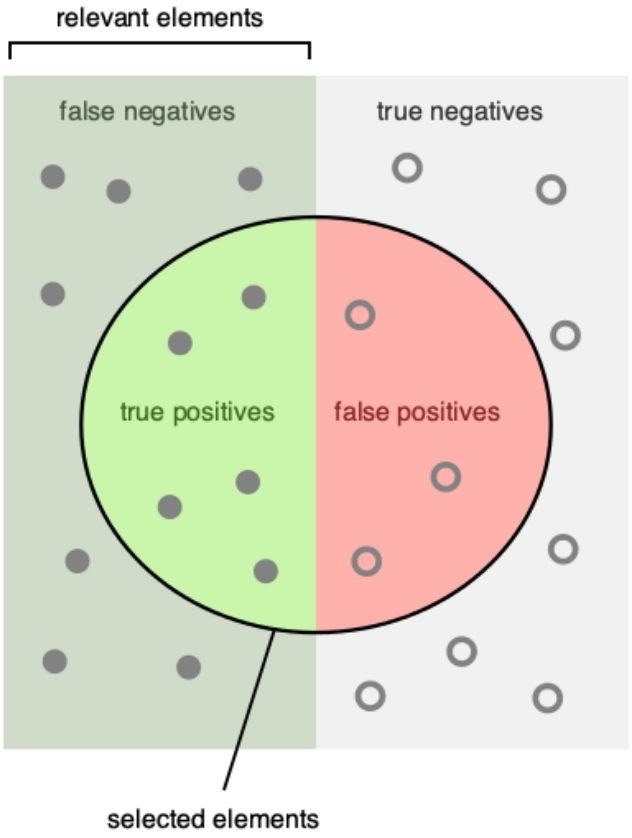
$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

  - **Recall**: fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Measuring quality of retrieval



$$P = tp/(tp+fp)$$
$$R = tp/(tp+fn)$$

$$F_{\beta=1} = \frac{2PR}{P+R}$$

F-measure is a weighted harmonic mean between *P* and *R*

Note: precision increases as recall decreases and vice versa; F-measure trades off *P* versus *R*

Measuring quality of retrieval

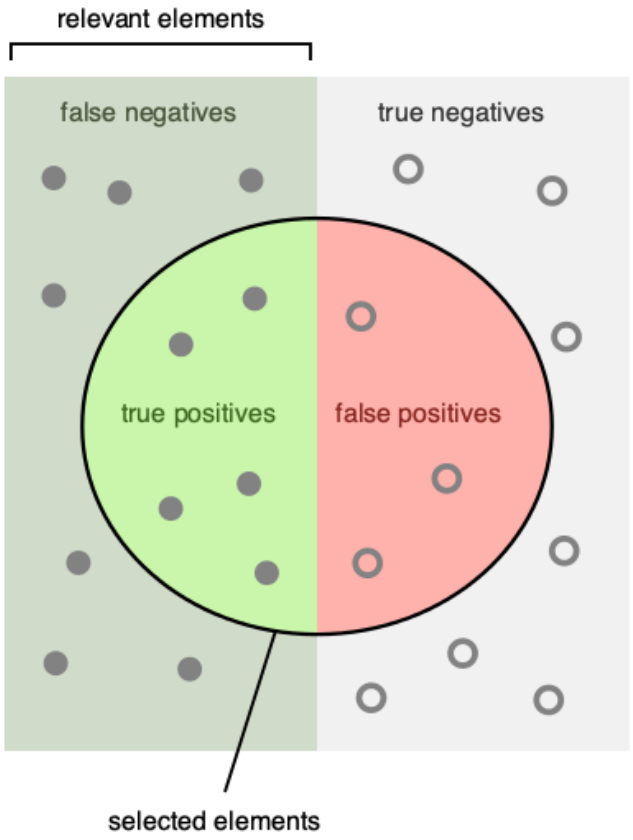
Example: An IR system returns 8 relevant documents, and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system **on this search**, and what is its recall? What is the F-measure?

- Hint: draw the contingency matrix

$$P = tp/(tp+fp)$$
$$R = tp/(tp+fn)$$

$$F_{\beta=1} = \frac{2PR}{P+R}$$

Measuring quality of retrieval



Sensitivity= $\frac{\text{true positives}}{\text{relevant elements}}$ = TP-rate = recall

Specificity = $\frac{\text{true negatives}}{\text{false positives} + \text{true negatives}}$ = TN-rate

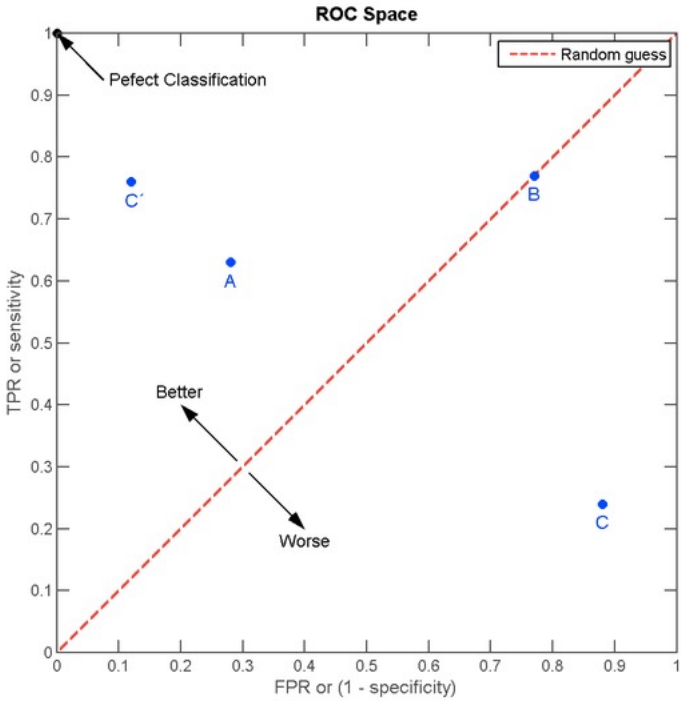
Specificity (**TN-rate**): usually huge TN (most documents are irrelevant), so this is not very informative for IR.

FP-rate: of all irrelevant documents, how many you wrongly predicted as positive.

Measuring quality of retrieval

- **ROC curve**
  - Plot TP-rate (sensitivity) against FP-rate for a series of queries

Relative trade-offs between TPs and FPs



https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Measuring quality of retrieval

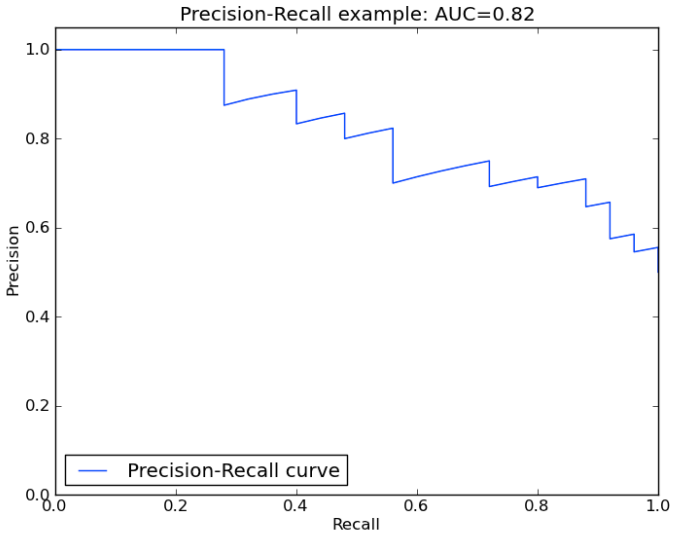
- *P, R* and *F* use unordered sets.
- What if we have ranked documents? The position should be taken into account.
- Several measures
  - precision-recall curve
  - precision at k
  - MAP
  - . . .
- Note: we looked at a single query so far - move to a set of queries.

https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

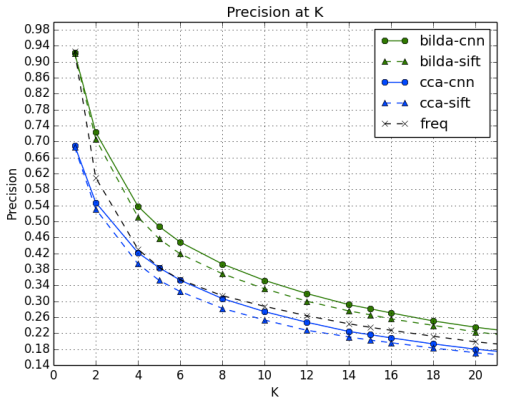
MANCHESTER 1824
The University of Manchester

Measuring quality of retrieval

- **Precision-recall curves**
  - At each (k+1)th step, if document k is relevant, than increase both precision and recall; otherwise, keep the recall, decrease the precision
  - Need to know number of relevant docs



- **Precision at k**
  - How many good results we have in the top k returned results
  - Doesn't need to know the total number of relevant docs
  - But not stable



https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Measuring quality of retrieval

- MAP = Mean Average Precision
  - Calculate average precision for each query and then find the mean over all queries
  - For each query, average precision is the average of precision values obtained for top k results each time a relevant document is retrieved

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

Average for query q_j

set of ranked retrieval results from the top result until you get to document d_k

https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Measuring quality of retrieval

- These previous measures (P, R, F, etc.) are often called **off-line** measures

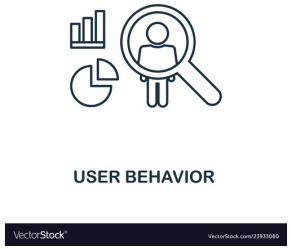
- **Online** metrics – based on user behaviour
  - User utility
  - Session abandon rate
  - Click-through rate
  - Etc.
  - *Note that these can be also used to change/influence ranking (see next Workshop)*

MANCHESTER 1824
The University of Manchester

User behavior

- User behavior is an intriguing source of relevance data
  - Users make (somewhat) informed choices when they interact with search engines
  - A lot of data available in search logs

- But there are significant caveats
  - User behavior data can be very noisy
  - Interpreting user behavior can be tricky
  - Spam can be a significant problem
  - Not all queries will have user behavior



Features based on user behavior

- **Click-through features**
  - Click frequency, click probability, click deviation
  - Click on next result? previous result? above? below?
- **Browsing features**
  - Cumulative and average time on page, on domain, on URL prefix; deviation from average times
  - Browse path features

[Agichtein, Brill, Dumais 2006; Joachims 2002]