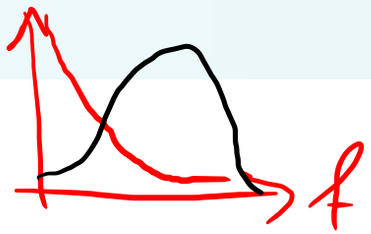
Lecture reviews — Week 08

J.-C. Chappelier & M. Rajman

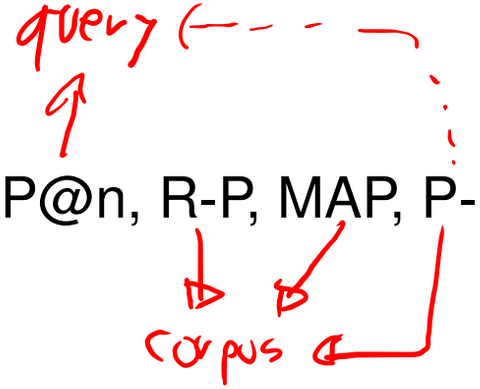
Laboratoire d'Intelligence Artificielle
Faculté I&C

Week 8 keypoints



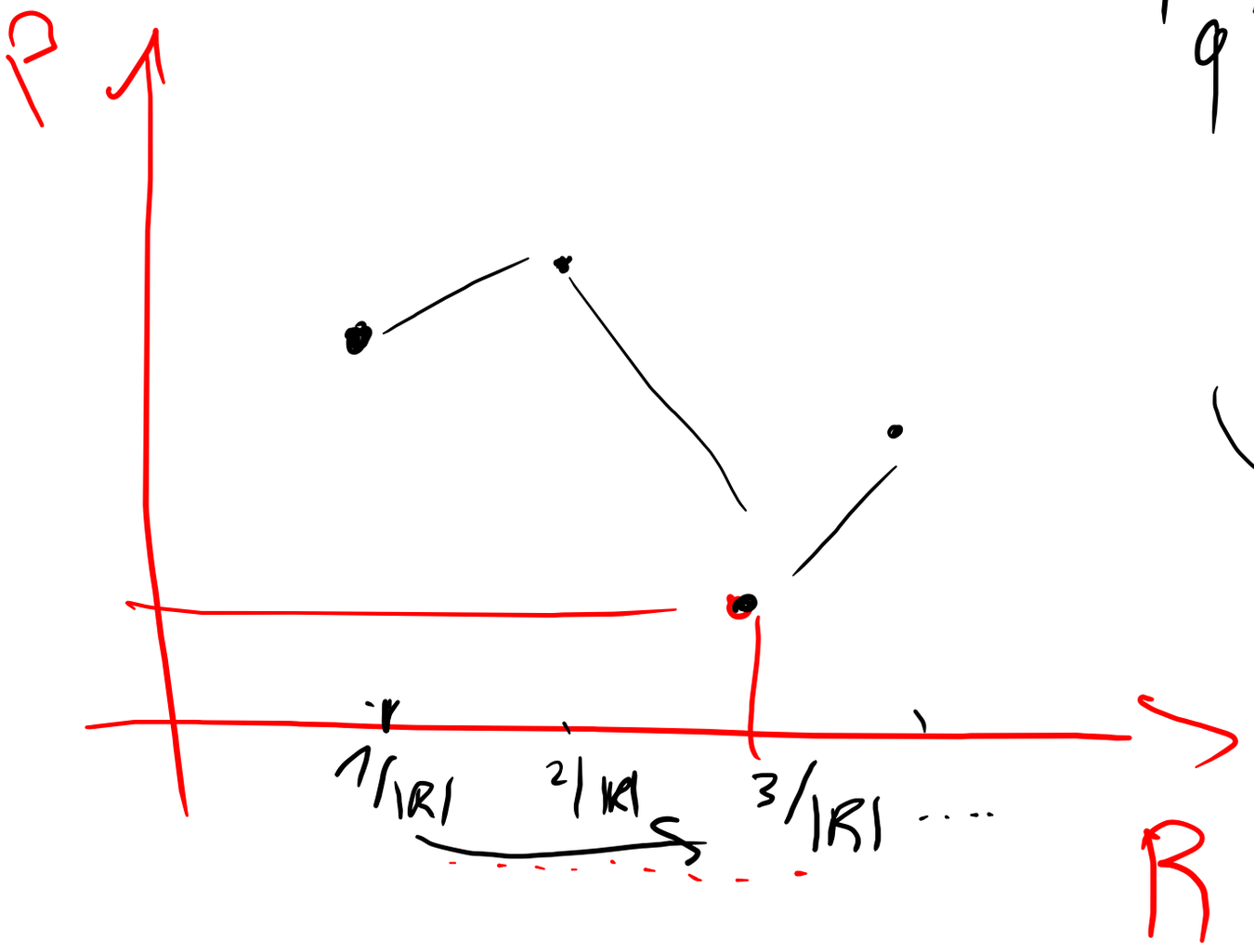
1
- preprocessing & indexing (tokenization, stemming/lemmatization, PoS-tag filtering, stop words, frequencies)
  (we could also add: sentence spliter, NERs, *n*-grams, parsers)
- weightings (desequentialisation): tf, tf-idf
- cosine similarity

2
- Information Retrieval (what, how)
- Information Retrieval evaluation metrics: P@n, R-P, MAP, P-R curves



3
- beyond standard vector space model:
  - topic models
  - word embeddings (and modern NLP)

J.-C. Chappelier & M. Rajman



query
q

corpus
set of doc

	relevant?
d_1	✓
d_2	×
⋮	✓
d_N	×

system:
return X doc
for X : 1 to $|R_q|$

Week 8 – study case 1

Using tf-idf weightning, what is the cosine similarity between these two "documents":

down fall fine wonder

Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her and to wonder what was going to happen next.

tf 1 1 1 1 → 1 2 1 2 : $\sqrt{10}$

Down, down, down. Would the fall never *come to an end? "I wonder how many miles I've fallen by this time?" she said aloud.*

tf 3 2 1 1 → 3 4 1 2 : $\sqrt{30}$

knowing that, for instance (invent your own if needed), among a corpus of 10'000 documents:

1'000 documents contain "*down*" 1
1'000 documents contain "*time*" 1
2 100 documents contain "*fall*"
2 100 documents contain "*wonder*"

idf → $\log \frac{|D|}{\#docs} = \log|D| - \log \#docs = 4 - \dots$

texts from "Alice's Adventures in Wonderland", Lewis Carroll (1865)

$$\cos(d_1, d_2) \propto \vec{d_1} \cdot \vec{d_2}$$

not 0 only on "$d_1 \cap d_2$"

$$\frac{3 \cdot 1 + 4 \cdot 2 + 1 \cdot 1 + 2 \cdot 2}{\sqrt{10}\ \sqrt{30}} = \frac{16}{\sqrt{10}\ \sqrt{30}}$$

Week 8 – study case 2

Compute R, P@5, R-prec, MAP and draw P-R curves for the two systems below

$|R_{q_1}| = 6$

query q_1

	system 1		system 2	
1	✔		✗	
2	✗		✔	
3	✗		✔	
4	✔		✔	
5	✔	3/5	✗	3/5
6	✗	3/6	✔	4/6
7	✗		✔	
8	✔		✔	
9	✗		✗	
10	✔	5/6	✗	1

$|R_{q_2}| = 7$

query q_2

system 1		system 2	
✗		✔	
✔		✔	
✔		✗	
✔		✗	
✔	4/5	✔	3/5
✔		✔	
✗	5/7	✗	4/7
✔		✔	
✗		✔	
✔	1	✔	1

$|R_{q_3}| = 8$

query q_3

system 1		system 2	
✔		✗	
✔		✗	
✔		✔	
✗		✔	
✔	4/5	✔	3/5
✔		✔	
✗		✔	
✔	6/8	✔	6/8
✗		✔	
✔	7/8	✔	1

knowing that, in the above results, for each query, at least one of the two systems retrieved all the relevant documents
(and assume the missing ones are retrieved at a very high rank)

$$P = \frac{\#\text{correctly retrieved}}{\#\text{system}}$$

$$R = \frac{\#\text{ correct. retrieved}}{\#\text{ claimed to correct}}$$

$$R\text{-Prec} = \frac{1}{\#\text{queries}} \sum_{q:\text{queries}} P_{@"R"}$$

$P_{@"R"}$ → @ rank where R = 1 COULD have been near $@|R_q|$

$$MAP = \frac{1}{\#\text{queries}} \sum_{q:\text{queries}} AP(q)$$

MAP: mean / average

$$AP(q) \hookrightarrow \frac{1}{|R_q|} \sum_{\text{ranks}} P_{@\text{rank}}$$

rank where system retrieved a R-doc