

Lecture reviews — Week 08 with solutions

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Week 8 keypoints

- ▶ preprocessing & indexing (tokenization, stemming/lemmatization, PoS-tag filtering, stop words, frequencies)
(we could also add: sentence splitter, NERs, n -grams, parsers)
- ▶ weightings (desequentialisation): tf, tf-idf
- ▶ cosine similarity
- ▶ Information Retrieval (what, how)
- ▶ Information Retrieval evaluation metrics: P@n, R-P, MAP, P-R curves
- ▶ beyond standard vector space model:
 - ▶ topic models
 - ▶ word embeddings (and modern NLP)

Week 8 – study case 1

Using tf-idf weighting, what is the cosine similarity between these two “documents”:

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.

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.

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

1’000 documents contain “*down*”

100 documents contain “*fall*”

1’000 documents contain “*time*”

100 documents contain “*wonder*”

texts from “Alice’s Adventures in Wonderland”, Lewis Carroll (1865)

Week 8 – study case (solution)

After some drastic filtering + normalisation, we could end-up with:

well deep fall slow time go down look wonder happen

down down down fall never come end wonder miles fall time say aloud

Interesting questions:

- ▶ *well* (noun): dropped by stop-list?
- ▶ *fell, fallen* → fall?
- ▶ *went* → go?
+keep it or stop list?
and what about “going”? “was going”?
- ▶ *slowly* → slow?
- ▶ *deep* → depth?
- ▶ keep *very? plenty? many? next? never?*

Week 8 – study case (solution)

Intersection of the two indexed documents:

down fall time wonder

down (3) fall (2) time wonder

idf: down: 1, fall: 2, time: 1, wonder: 2

$d_1: (1 \times 1, \quad 1 \times 2, \quad 1 \times 1, \quad 1 \times 2), \quad \text{norm: } \sqrt{10}$

$d_2: (3 \times 1, \quad 2 \times 2, \quad 1 \times 1, \quad 1 \times 2), \quad \text{norm: } \sqrt{30}$

$$\cos(d_1, d_2) = \frac{16}{\sqrt{10}\sqrt{30}} \simeq 0.9238$$

Week 8 – study case 2

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

query q_1			query q_2			query q_3		
	system 1	system 2		system 1	system 2		system 1	system 2
1	✓	✗		✗	✓		✓	✗
2	✗	✓		✓	✓		✓	✗
3	✗	✓		✓	✗		✓	✓
4	✓	✓		✓	✗		✗	✓
5	✓	✗		✓	✓		✓	✓
6	✗	✓		✓	✓		✓	✓
7	✗	✓		✗	✗		✗	✓
8	✓	✓		✓	✓		✓	✓
9	✗	✗		✗	✓		✗	✓
10	✓	✗		✓	✓		✓	✓

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)

Week 8 – study case 2 (solution)

query q_1					
system 1			system 2		
P	R		P	R	
✓	1/1	1/6	✗	0/1	0/6
✗	1/2	1/6	✓	1/2	1/6
✗	1/3	1/6	✓	2/3	2/6
✓	2/4	2/6	✓	3/4	3/6
✓	3/5	3/6	✗	3/5	3/6
✗	3/6	3/6	✓	4/6	4/6
✗	3/7	3/6	✓	5/7	5/6
✓	4/8	4/6	✓	6/8	6/6
✗	4/9	4/6	✗	6/9	6/6
✓	5/10	5/6	✗	6/10	6/6
R-P 3/6			4/6		
AvgP 0.52			0.67		

query q_2					
system 1			system 2		
P	R		P	R	
✗	0/1	0/7	✓	1/1	1/7
✓	1/2	1/7	✓	2/2	2/7
✓	2/3	2/7	✗	2/3	2/7
✓	3/4	3/7	✗	2/4	2/7
✓	4/5	4/7	✓	3/5	3/7
✓	5/6	5/7	✓	4/6	4/7
✗	5/7	5/7	✗	4/7	4/7
✓	6/8	6/7	✓	5/8	5/7
✗	6/9	6/7	✓	6/9	6/7
✓	7/10	7/7	✓	7/10	7/7
5/7			4/7		
0.71			0.75		

query q_3					
system 1			system 2		
P	R		P	R	
✓	1/1	1/8	✗	0/1	0/8
✓	2/2	2/8	✗	0/2	0/8
✓	3/3	3/8	✓	1/3	1/8
✗	3/4	3/8	✓	2/4	2/8
✓	4/5	4/8	✓	3/5	3/8
✓	5/6	5/8	✓	4/6	4/8
✗	5/7	5/8	✓	5/7	5/8
✓	6/8	6/8	✓	6/8	6/8
✗	6/9	6/8	✓	7/9	7/8
✓	7/10	7/8	✓	8/10	8/8
6/8			6/8		
0.76			0.64		

$$\text{R-Prec}_1 = \frac{1}{3} \left(\frac{3}{6} + \frac{5}{7} + \frac{6}{8} \right) \simeq 0.65$$

$$\text{AvgP}_1(q_1) = \frac{1}{6} \left(1 + \frac{2}{4} + \frac{3}{5} + \frac{4}{8} + \frac{5}{10} \right) \simeq 0.517$$

$$\text{MAP}_1 \simeq 0.66$$

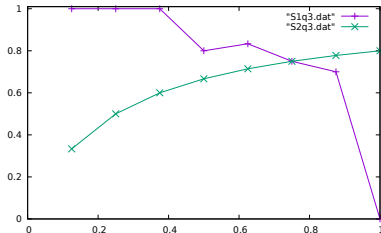
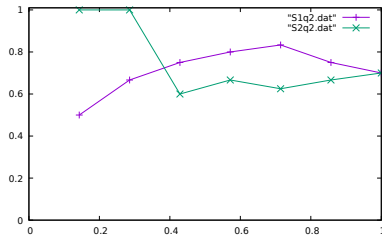
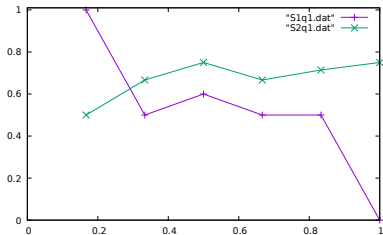
$$\text{R-Prec}_2 = \frac{1}{3} \left(\frac{4}{6} + \frac{4}{7} + \frac{6}{8} \right) \simeq 0.66$$

$$\text{AvgP}_2(q_1) = \frac{1}{6} \left(\frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \frac{4}{6} + \frac{5}{7} + \frac{6}{8} \right) \simeq 0.675$$

$$\text{MAP}_2 \simeq 0.69$$

Week 8 – study case 2 (solution)

raw plots for each query:



in practice: some averaging and some filtering (keep only max Prec) is done to ensure monotonically decreasing (or constant) curves