Week 8 keypoints

- 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

  word

  (word, doc)
- cosine similarity
- Information Retrieval (what, how)

  (q) (q)
- Information Retrieval evaluation metrics: P@n, R-P, MAP, P-R curves

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

$$MAP = \underset{q}{\text{mean}} \; AP(q)$$

$$AP(q) = \frac{1}{|R(q)|} \sum_{d \in R(q)} P@\text{rank}(d)$$

q: $R(q)$ set of relevant docs for q

System answer to q

d_1

$\boxed{d_2}$ $P@2$

$\boxed{d_3}$ $P@3$

d_4

$\boxed{d_5}$ $P@5$

Relevant

Week 8 – study case 1

	down	time	fall	wonder	...
d_1	1 · 1	1 · 1	1 · 2	1 · 2	0 0 0
d_2	3 · 1	1 · 1	2 · 2	1 · 2	000

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

d_1 *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.*

→ fall

can

d_2 *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*" → $\log_{10}\left(\frac{10'000}{1000}\right) = 1$

1'000 documents contain "*time*" → 1

100 documents contain "*fall*" → 2

100 documents contain "*wonder*" → 2

$$idf = \log\left(\frac{|D|}{|doc \ni word|}\right)$$

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

J.-C. Chappelier & M. Rajman

Week 8 – study case 2

$R = \frac{\text{nb rel doc retreived}}{\text{nb rel. doc.}}$

Compute R, P@5, R-prec, MAP and draw P-R curve(s) for the two systems below

$|R(q_1)| = 6$ $|R(q_2)| = 7$ $|R(q_3)| = 8$

	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	✓	✗	✓	✓	✓	✓

$R = \frac{5}{6}$ $\underline{}\ 5$ $\underline{}\ 6 \rightarrow R < 1$ $R = 1$ $R = 1$ $R = \frac{7}{8}$ $R : 1$

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)

Week 8 – study case 2

→ overall : average over q_i : $\frac{1}{3}\sum_{i=1}^{3} \boxed{?}$

$X(q)$

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

for Syst 1:

$|R(q_1)| = 6$ 7 8

	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	✓	✗	✓	✓	✓	✓

$X(q_1) = P@6$ $\frac{3}{6} = \frac{1}{2}$

$X(q_2) = P@7$ $\frac{5}{7}$

$X(q_3) = P@8$ $\frac{6}{8}$

↳ R-Prec for Syst 1 = $\frac{1}{3}\left(\frac{1}{2} + 5/7 + 3/4\right)$

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

→ overall q: $\frac{1}{3}\left(AP(q_1) + AP(q_2) + AP(q_3)\right)$

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

$AP(q_1) = \frac{1}{6}\left(P@1 + P@4 + P@5 + P@8 + P@10 + 0\right)$

$|R(q_1)| = 6$ $\qquad \frac{1}{7}(\dots)$ $\quad 7$ $\qquad \frac{1}{8}(\dots)$ $\quad 8$

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

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

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

	query q_1		query q_2 7		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	✔	✗	✔	✔	✔	✔

R:

← fix

$\frac{1}{7}$

2/7

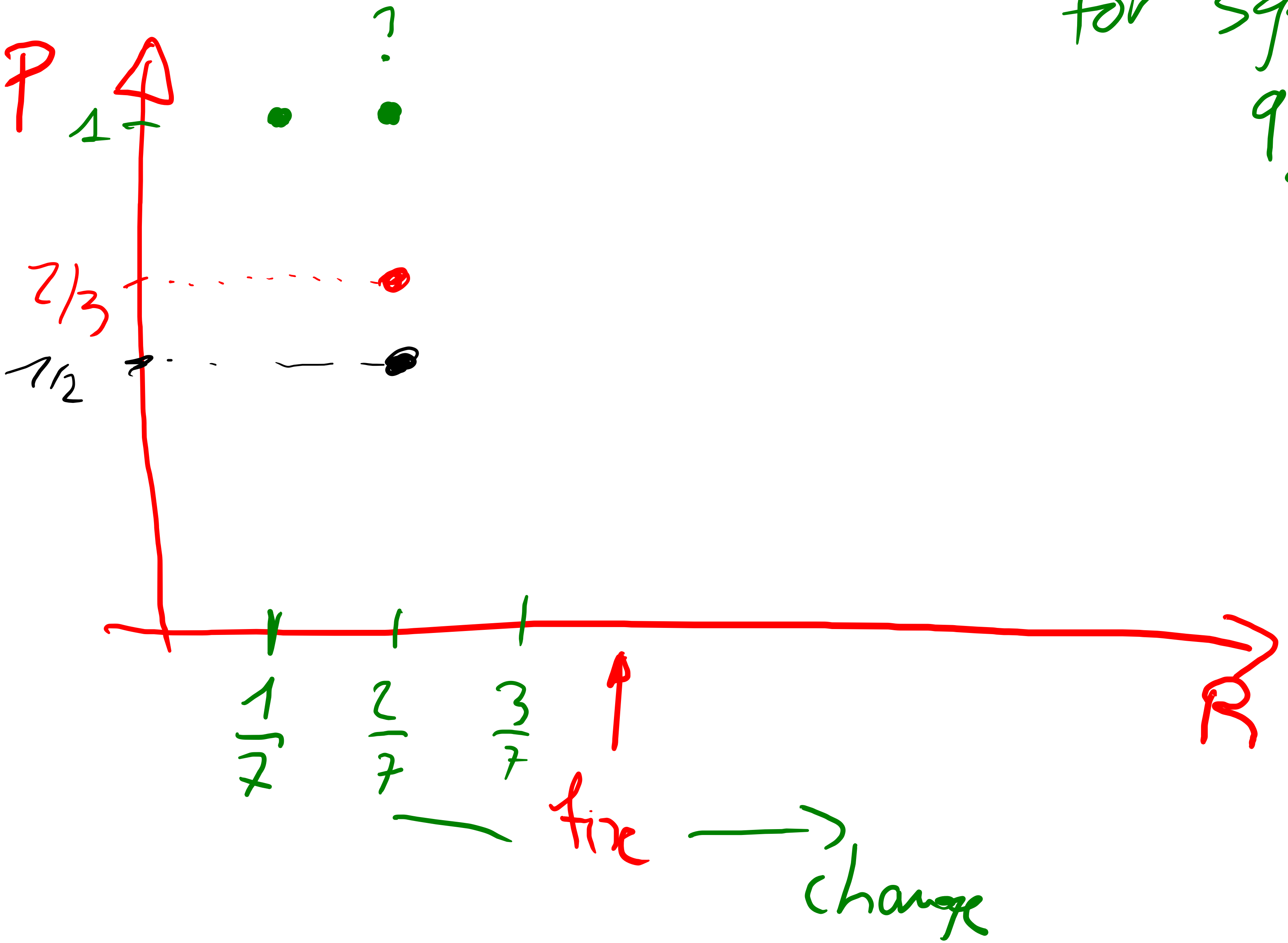
2/7

2/7

change

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)

for syst 2 q_2


P
2/3
1/2
?
1/7
2/7
3/7
R
fix
change