

P3 discussion

See last 30 minutes or so of the recordings for
April 25 and April 27

P3 retrieve

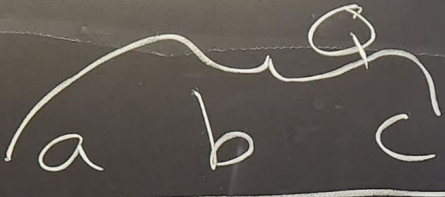
① indexes SciAnn

(Ch.5) JSON - StoryID
Text — tokenized
(URL) stemmed

② runs queries

(Ch.7) AND $w_{p_1} \wedge w_{p_2} \wedge w_{p_3} \dots$
OR
BM25
QL (Dir.)

↑ if contains space
then is a phrase

OR  // Given idx, the index
// Returns list of story IDs

results = Set that is empty

for q in Q

Set storiesWith_ q = idx.getStories(q)

results.add(storiesWith_ q)

return results

with term or phrase

AND

results = getStoner($Q[0]$)

for rest of q in Q
nextSet
get q 's ~~stay~~ IDs

BM 25

avg d.l
doc length
tf
ctf

for each r in results

if r is not in nextSet
remove r from results

OR a b c ^Q // Given idx, the index
 // Returns list of stay IDs

results = Set that is empty

for q in Q

Set stonesWith-q = idx, getStones(q)

results.add(stonesWith-q)

return results

with term or phrase

$$\sum_{q \in Q} \log \frac{QLDir(doc)}{(1-\alpha) \frac{tf_q}{idf_q} + \alpha}$$

results = getStones(Q[0])

for rest of q in Q

nextSet = get q's ~~next~~ stay IDs

for each r in results

if r is not in nextSet
 remove r from results

BM25

avg doc length

tf
 idf

P3 retrieve

① indexes SciAnn

(Ch.5) JSON - StoryID
Text — tokenized
(URL) — stemmed

② runs queries

(Ch.7) AND $w_{p_1} \wedge w_{p_2} \wedge w_{p_3} \dots$
OR
BM25
QL (Dir.)

↑ if contains space
then is a phrase

Need from idx

list of docs with a term

list of docs with a phrase

TF (term/phrase) — from inverted list

CTF (") — store while indexing

Doclen (d) — Map inside idx

numTermOccurrences()

avgdl

(42, 106, 107, ...)

(42:3, 106:102, ...)

(42:3 [6, 12, 107],

106:102 [-, -])

QL // return a <story, score> tuples

for docid : range(numDocs) $\in [1, \text{numDocs}]$

score = 0

for q in Q

tf = ~~idx~~ idx.getTF(q, docid)

doclen = idx.doclen(docid)

ctf = idx.getCTF(q)

numTermOccurrences = idx.numTerms(q)

score += f($\frac{tf \cdot ctf}{\sqrt{\text{numTermOccurrences}}}$)

if score > 0
add to results

end