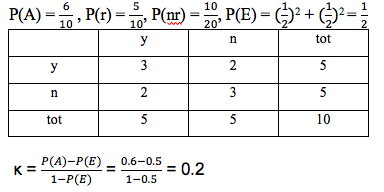
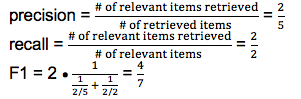
Collaborators: Ben Hlewka, Calvin Lee, Cam McIntrye  
Consultations: Class slides, An Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze

1.

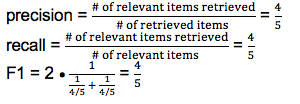
a)



b)



c)



2.

a)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | q0 | d1 | d2 | d3 | Qr | Qnr | qm |
| i’ve | 0 | 1 | 0 | 0 | 2 | 0 | 0.70 |
| seen | 0 | 1 | 1 | 0 | 2 | 1 | 0.55 |
| fire | 1 | 1 | 0 | 1 | 2 | 0 | 1.2 |
| and | 0 | 1 | 0 | 0 | 1 | 0 | 0.35 |
| rain | 1 | 1 | 1 | 0 | 1 | 1 | 0.7 |
| have | 0 | 0 | 1 | 0 | 0 | 1 | 0 (-0.35) |
| you | 0 | 0 | 1 | 0 | 0 | 1 | 0 (-0.35) |
| ever | 0 | 0 | 1 | 0 | 0 | 1 | 0 (-0.35) |
| the | 0 | 0 | 1 | 0 | 0 | 1 | 0 (-0.35) |
| she | 0 | 0 | 0 | 1 | 1 | 0 | 0 (-0.35) |
| ran | 0 | 0 | 0 | 1 | 1 | 0 | 0.35 |
| calling | 0 | 0 | 0 | 1 | 1 | 0 | 0.35 |
| wild | 0 | 0 | 0 | 1 | 1 | 0 | 0.35 |

Negative qm replaced with 0’s

qm = alpha · q0 + beta · Dr - gamma · Dnr

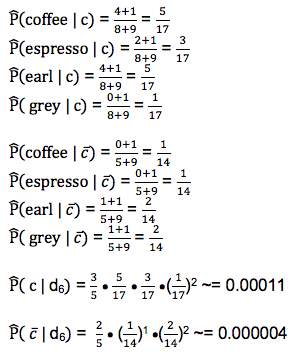
qm = 0.5 · q0 + 0.35 · d1 \* d3 – 0.15 · d2

Therefore, a revised query could be “fire”, as it would return documents 1 and 3, while not returning document 2.

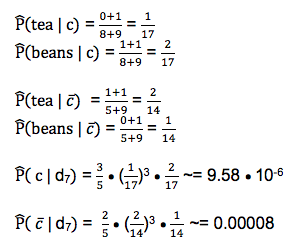
b) Since relevance feedback relies on user feedback, it may increase precision for some users, while decreasing it for others. In general, relevance feedback will increase recall, but it will not necessarily increase precision for users. Most users will also just be searching for something, not wanting to provide feedback on the queries and results.

3.

Test1:



Test 2:



docID 6: Yes

docID 7: No

4.

a)

M1: 0.040 · 0.051 · 0.114 · 0.236 · 0.005 · 0.003 · 0.0003 = 2.5 · 10^-13

M2: 0.051 · 0.02 · 0.114 · 0.236 · 0.005 · 0.003 · 0.0003 = 1.2 · 10^-13

b)

M1: 0.003 · 0.02 · 0.137 · 0.102 · 0.001 · 0.018 · 0.0002 = 3.0 · 10^-15

M2: 0.047 · 0.003 · 0.137 · 0.102 · 0.001 · 0.018 · 0.0002 = 7.0 · 10^-15

c)

Using a bigram model instead will improve probabilities. Since words are now being related to another eg. taylor swift vs. taylor, in pairs, probabilities for each term will be better reflected.

5.

a)

tfidf = (1+ logtf)(log(N/df))

Cluster c1:

bob: (1 + log1)(log3) = 0.47712

dylan: 0.47712

is: (1+ log3)(log(3/3)) = 0

a: 0

folk: 0

singer: 0

neil: 0.47712

diamond: 0.47712

james: 0.47712

taylor: 0.47712

and: 0.47712

pop: 0.47712

rock: 0.47712

d1: 0.95424, d2: 1.908, d3: 1.43136

μc1 = ⅓ · ( 0.95424 + 1.908 + 1.43136) = 1.4312

Cluster c2:

fleetwood: 0.30103

mac: 0.30103

is: 0

a: 0

rock: 0

chicago: 0.30103

and: 0.30103

blues: 0.30103

group: 0

d1: 1.20412, d2: 0.30103

μc2 = 1/2 · ( 1.20412 + 0.30103) = 0.752575

Cluster c3:

stranger: 0.30103

things: 0.30103

a: 0

netflix: 0.30103

science: 0.30103

fiction: 0.30103

supernatural: 0.30103

series: 0

chronicles: 0.30103

the: 0

mysterious: 0.30103

happenings: 0.30103

in: 0.30103

hawkins: 0.30103

indiana: 0.30103

freaks: 0.30103

and: 0.39165

geeks: 0.30103

a: 0

series: 0

which: 0.30103

was: 0.30103

cancelled: 0.30103

way: 0.30103

too: 0.30103

soon: 0.30103

follows: 0.30103

the: 0

weir: 0.30103

children: 0.30103

their: 0.30103

groups: 0.30103

of: 0.30103

friends: 0.30103

d1 = 3.61236, d2 = 5.20813

μc3 = 1/2 · ( 3.61236 + 5.20813) = 4.410245

b)

Query: is a folk

Cluster 1 would be returned as all documents 1, 2, and 3 contain the query.

c)

Query: rock

Cluster 2 contains documents 4 and 5, which contain the query, but document 3, which also contains “rock”, would be missed from the query.