1a) To Compute their RSV for the doc 11, doc 12

Let analyze the relevance of each term in the query vector using document-term Matrix

Summary of T1 to T6 in terms of their occurrence in relevant and no relevant documents.

Term	Relevance (R	Non-Relevance(NR)
T1	3	3
T2	3	2
T3	2	3
T4	3	4
T5	1	5
T6	3	1

Given query vector (1, 1, 0, 1, 0, 1) and 0.5 to Smooth Counts

S = number of documents containing the term that are relevant.

S = total number of relevant documents in the doc-term-matrix

n-s = number of documents not Containing the term that are relevant.

N-S = total number of non-relevant documents rows

N = total number of documents = 10

$x_i = 1$ if the term is in the document, $x_i = 0$ if

For each term in the query, we compute $P(R|x_i)$ where $x_i = 1$ and then multiply these probabilities together

RSV for doc11 => (0, 1, 1, 0, 0, 1) and doc2 => (1, 0, 1, 1, 0, 1)

For Doc 11

T1 = 0 : $S-s+0.5$ and $N-n-S+s+0.5$

T2 = 1 : $s+0.5$ and $n-s+0.5$

T3 = 1 : $s+0.5$ and $n-s+0.5$

T3 = 1 : s+0.5 and n-s+0.5

T4 = 0 : S-s+0.5 and N-n-S+s+0.5

T5 = 0 S-s+0.5 and N-n-S+s+0.5

T6 = 1 : S+0.5 and n-s+0.5

For doc 12

T1 = 1 : s+0.5 and n-s+0.5

T2 = 0 : S-s+0.5 and N-n-S+s+0.5

T3 = 1 : s+0.5 and n-s+0.5

T4 = 1 : s+0.5 and n-s+0.5

T5 = 0 : S-s+0.5 and N-n-S+s+0.5

T6 = 1 : s+0.5 and n-s+0.5

For a term i where $x_i = 1$

$$RSV = \log\left(\frac{s+0.5}{n-s+0.5}\right) + \log\left(\frac{N-n-S+s+0.5}{S-s+0.5}\right)$$

$x_i = 0$

$$RSV = \log\left(\frac{S-s+0.5}{N-n-S+s+0.5}\right) + \log\left(\frac{n-s+0.5}{s+0.5}\right)$$

For doc 11 (T1=0) S=3 for T1 S=4 n

(NR+R) ⇒ n = total number of documents where the term appears = 6

N = Total document (N) = 10

For doc 11 (T1=0)

$$= \log\left(\frac{4-3+0.5}{10-6-4+3+0.5}\right) + \log\left(\frac{6-3+0.5}{3+0.5}\right)$$

$$= \log(1.5/3.5) + \log(1)$$

$$\approx -0.85$$

Using base 10

For doc 12 (T1 = 1)

$$= \log\left(\frac{3+0.5}{6-3+0.5}\right) + \log\left(\frac{10-6-4+3+0.5}{4-3+0.5}\right)$$

$$\approx 0.37$$

Doc 12 T1 has a higher RSV contribution (1a)

(T2) For doc s=3, S=4, n=5, N=10

11 (T2=1) $= \log\left(\frac{3+0.5}{5-3+0.5}\right) + \log\left(\frac{10-5-4+3+0.5}{4-3+0.5}\right) \approx 0.62$

$\rightarrow 0.44 + 0.48$

For doc 12 (T2 = 0)

$= \log\left(\frac{4-3+0.5}{10-5-4+3+0.5}\right) + \log\left(\frac{5-3+0.5}{3+0.5}\right) \approx 0.3$

$-0.48 \quad +0.71$

(T3) Doc 11 T2 has a higher RSV contribution

For doc 11 for (T3) s=2, S=4, n=5 N=10

Dc 11 T3=1 and Doc 12 T3=1 has equal RSV contribution

(T4)

for doc 11 for (T4) s=3, S=4, n=7, N=10

T4=0 $= \log\left(\frac{4-3+0.5}{10-7+4+0.5}\right) + \log\left(\frac{7-3+0.5}{3+0.5}\right) \approx -0.11$

$+3 \qquad -0.22$

For doc 12, T4 = 1

$= \log\left(\frac{3+0.5}{7-3+0.5}\right) + \log\left(\frac{10-7-4+3+0.5}{4-3+0.5}\right) \approx 0.11$

-0.11

Doc 12 T4 has a higher RSV contribution

(T5) For doc 11 T5=0, doc 12 T5=0
s=1, S=4, n=6, N=10 they will have equal RSV contribution

(T6) For doc 11 T6=1, doc 12 T6=1
s=3, S=4, n=4, N=10 they have equal RSV contribution

doc 11 RSV
$= -0.85 + 0.62 + 0.11 = -0.34$

doc 12 RSV
$= 0.37 + 0.3 + 0.11 = 0.78$

doc 12 RSV is higher doc 11

10. using

$$BM25 = \sum_{t \in d} \left(IDF_{(t)} \times \frac{(K_1+1)\cdot tf_i}{tf_i + \left[(1-b)+b * \frac{dl}{avgdl}\right] * K_1} \right.$$

$IDF_{(t)} \Rightarrow \log \frac{N}{df_i} \Rightarrow$ inverse document frequency of term t

$tf_i \Rightarrow$ term frequency of t in document d

$K_1 \Rightarrow$ 2. $b = 0.75$, $avgdl = 1600$

$|d_1| = 1000$, $|d_2| = 1,500$, $|d_3| = 1,200$

IDF (Car) = 1.65, IDF (Auto) = 2.08, IDF (Insurance) = 1.62

IDF (Best) = 1.5

To Calculate BM25 score for each documents

$$BM25(d_1) = \sum \left\{ \begin{array}{l} 1.65 \times \frac{(2+1)\times 27}{27 + 2 \times \left[(1-0.75)+0.75 \times \frac{1000}{1600}\right]} + \\ 2.08 \times \frac{(2+1)\times 3}{3 + 2 \times \left[(1-0.75)+0.75 \times \frac{1000}{1600}\right]} + \\ 1.62 \times \frac{(2+1)\times 0}{0 + 2 \times \left[(1-0.75)+0.75 \times \frac{1000}{1600}\right]} + \\ 1.5 \times \frac{(2+1)\times 14}{14 + 2 \times \left[(1-0.75)+0.75 \times \frac{1000}{1600}\right]} \end{array} \right.$$

xxx

$$BM25(d_2) \sum \left[\begin{array}{l} 1.65 \times \frac{(2+1)\times 4}{4 + 2 \times \left[(1-0.75)+0.75 \times \frac{1500}{1600}\right]} + \\ 2.08 \times \frac{(2+1)\times 33}{33 + 2 \times \left[(1-0.75)+0.75 \times \frac{1500}{1600}\right]} + \\ 1.62 \times \frac{(2+1)\times 33}{33 + 2 \times \left[(1-0.75)+0.75 \times \frac{1500}{1600}\right]} + \end{array} \right.$$

$$BM25(d_2) \left[1.5 \times \frac{(2+1)\times 0}{0+2\times\left[(1-0.75)\times 0.75 \times \frac{1500}{1600}\right.} \right]$$

$$BM25(d_3) \left\{ \begin{array}{l} 1.65 \times \frac{(2+1)\times 24}{24+2\times\left[(1-0.75)\times 0.75\times \frac{1200}{1600}\right.} + \\ 2.08 \times \frac{(2+1)\times 0}{0+2\times\left[(1-0.75)\times 0.75\times \frac{1200}{1600}\right]} + \\ 1.62 \times \frac{(2+1)\times 29}{29+2\times\left[(1-0.75)\times 0.75\times \frac{1200}{1600}\right]} + \\ 1.5 \times \frac{(2+1)\times 17}{17+2\times\left[(1-0.75)\times 0.75\times \frac{1200}{1600}\right]} \end{array} \right]$$

2) d_1: click go the shears boys click click click
d_2: Click click
d_3: metal here
d_4: metal shears click here

$\lambda = 0.5$

the model will be MLE unigram model from the documents and Collection, mixed.

Query 1 : click

$P(q|d_1) = \lambda 4/8 + 1-\lambda 7/16 = 0.5 \times 1/2 + 0.5 \times 7/16 \simeq 0.47$

$P(q|d_2) = 0.5 \times 1 + 0.5 \times 7/16 \simeq 0.719$

$P(q|d_3) = 0.5 \times 0 + 0.5 \times 7/16 \simeq 0.22$

$P(q|d_4) = 0.5 \times 1/4 + 0.5 \times 7/16 \simeq 0.345$

So the ranking is $d_2 > d_1 > d_3 > d_4$

query 2 Shears

$P(q|d_1) = 0.5 \times 1/8 + 0.5 \times 2/16 = 0.063 + 0.063 = 0.126$

$P(q|d_2) = 0.5 \times 0 + 0.063 = 0.063$

$P(q|d_3) = 0.5 \times 0 + 0.063 = 0.063$

$P(q|d_4) = 0.5 \times 1/8 + 0.5 \times 2/16 = 0.126$

Ranking $\Rightarrow$ $d_1, d_4 \rightarrow d_2, d_3$

query 3 click shears

$P(q|d_1) = \left[0.5 \times 4/8 + 0.5 \times 7/16\right] * \left[0.5 \times 1/8 + 0.5 \times 2/16\right] = 0.47 + 0.126 = 0.5$

$P(q|d_2) = 0.719 + 0.063 = 0.78$

$P(q|d_3) = 0.22 + 0.063 = 0.283$

$P(q|d_4) = 0.345 + 0.126 = 0.471$

Ranking $d_2 > d_1 > d_4 > d_3$

3.) Evaluation of Unranked list

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#\text{ retrieved items}} = P(\text{relevant} \mid \text{retrieved})$$

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{total relevant documents})} = P(\text{retrieved} \mid \text{relevant in Coll})$$

These notion can be made clear by examing the following Contigency Table

	Relevant	Not Relevent
Retrieved	TP	FP
Not Retrieved	FN	TN

Total_retrieved_docs = 20, Total_relevant_items = 8
relevant_items_retrieved = 6

a. Precision of the system on the top 20 = $6/20 = 0.3$

b. F1 score on the top20 = $\frac{2 \times (\text{Precision\_at\_20} \times \text{recall\_at\_20})}{(\text{Precision\_at\_20} \times \text{recall\_at\_20})}$

To calculate Recall_at_20 = $6/8 = 0.75$

$$= \frac{2 \times (0.3 \times 0.75)}{0.3 + 0.75} = 0.428$$

c. Uninterpolated precision at 25%, where requred_recall = 0.25

To calculate this we created a table to analyze precision and recall at each retrieved items.

i	item	Precision_at_i	Recall_at_i	Cumulative count of Relevant
1	R	1/1 = 1	1/8 → 0.125	1
2	R	2/2 = 1	2/8 → 0.25	2
3	N	2/3 → 0.67	0.25	
4	N	2/4 → 0.5	0.25	
5	N	2/5 → 0.4	0.25	
6	N	2/6 → 0.33	0.25	
7	N	2/7 → 0.286	0.25	
8	N	2/8 → 0.25	0.25	
9	R	3/9 → 0.33	0.375	3
10	N	3/10 → 0.3	0.375	
11	R	4/11 → 0.36	0.5	4
12	N	4/12 → 0.33	0.5	
13	N	4/13 → 0.31	0.5	
14	N	4/14 → 0.29	0.5	
15	R	5/15 → 0.33	0.625	5
16	N	5/16 → 0.31	0.625	
17	N	5/17 → 0.29	0.625	
18	N	5/18 → 0.28	0.625	
19	N	5/19 → 0.26	0.625	
20	R	6/20 → 0.3	0.75	6

c. Uninterpolated precision at 25% recall = 1 at first value which recall >= 25%

d. Interpolated precision at 33% (37.5%) _recall : This will be taking the maximum of precision >= 33% of recall

max [0.33, 0.3, 0.36, 0.33, 0.31, 0.29, 0.33, 0.31, 0.28, 0.26, 0.3]

= 0.36

e. Mean Average Precision MAP for the query

We Isolate Precision based on relevant item

= 0.125 + 0.25 + 0.375 + 0.5 + 0.625 + 0.75

3) e $\quad map\_for\_query = Sum(average\_Precisions) / total\_relevant\_doc_{in\ Coll}$

$= 1 + 1 + 0.33 + 0.36 + 0.33 + 0.3 = 3.32$

$= \cancel{3.32/20} \quad 3.32/8 \cong 0.415 \Rightarrow 41.5\%$

3f) To achieve the largest possible MAP that this system could have? - This could be achieve if all the relevant document were ranked at the top of the retrieved list

$\Rightarrow 1+1+1+1+1+1+1+1 = 6/8 = 75\%$

3g) For Smallest MAP this could achieved ~~be achieved~~ by ranking the relevant art the end of the list

$\Rightarrow 1/15 + 2/16 + 3/17 + 4/18 + 5/19 + 6/20 \Rightarrow$

$$= \frac{0.067 + 0.125 + 0.18 + 0.22 + 0.263 + 0.3}{8}$$

$= 1.155/8 = 0.144 \Rightarrow 14.42\%$

3h) - The error from the estimated MAP to the largest possible MAP is: $0.75 - 0.414 = 0.336$

The error from the estimated MAP to the Smallest possible MAP: $0.414 - 0.1442 = 0.2698$

$(0.336, 0.2698)$