Paper Code	Examiner	Department	Office
INT309	Fangyu Wu	Department of Intelligence Science	SD555

Question 1

Suppose we have a collection of 20 documents, $d_1, d_2, ..., d_{20}$, which have been judged for relevance to a query. A 3-point relevance scale was used, so relevant documents have been divided into Perfect, Good and just Relevant results. Weights for these levels are shown below:

Perfect	3
Good	2
Relevant	1
Non-relevant	0

Here are the documents and their judgments:

$$\begin{aligned} \text{Perfect} &= \{d_1, d_8\} \\ \text{Good} &= \{d_4, d_9\} \\ \text{Relevant} &= \{d_2, d_7, d_{10}\} \\ \text{Non-relevant} &= \{d_3, d_5, d_6, d_{11} - d_{20}\} \end{aligned}$$

Relevant

Consider now these two ordered result lists retrieved by different systems:

$$\begin{aligned} Result_1 &= < d_1, d_2, d_7, d_3, d_4, d_8, d_6, d_5 > \\ Result_2 &= < d_1, d_4, d_8, d_2, d_6, d_3, d_5, d_7 > \end{aligned}$$

1. What are the precision and recall for result list $Result_2$?
2. What is the precision @4 of each result list?
3. What is the average precision of each result list?
4. To measure/evaluate information retrieval (IR) effectiveness, what are the three elements required for a test collection, so the performance of the IR system could be compared?

Question 1

1. Result2 = $< d_1, d_4, d_8, d_2, d_6, d_3, d_5, d_7 >$

Precision = $\frac{5}{8} = 62.5\%$

Relevant files are $d_1, d_2, d_4, d_7, d_8, d_9, d_{10}$

$\therefore$ Recall = $\frac{5}{7} \approx 71.4\%$

2. Result 1: Precision@4 $= \frac{3}{4} = 75\%$

Result 2: Precision@4 $= \frac{4}{4} = 100\%$

3. Result 1:

$\langle d_1 \checkmark\ d_2 \checkmark\ d_7 \checkmark\ d_3\ d_4 \checkmark\ d_8 \checkmark\ d_6\ d_5 \rangle$

P: 100% 100% 100% 75% 80% 83% 71% 62.5%

$\text{AvgP\_1} = \frac{1+1+1+0.8+0.83}{5} \approx 92.6\%$

$\langle d_1 \checkmark, d_4 \checkmark, d_8 \checkmark, d_2 \checkmark, d_6, d_3, d_5, d_7 \checkmark \rangle$

P: 100% 100% 100% 100% - - - - 62.5%

$\text{AvgP\_2} = \frac{1+1+1+1+62.5\%}{5} = 92.5\%$

4.
- 1) benchmark document collection
- 2) benchmark suite of information needs expressible as queries
- 3) an assessment of either relevant/irrelevant judgements for each query-doc pair

Question 2

Consider following documents with the stop word list: [when, in, the, and, I]
Doc 1: when walking in the rain
Doc 2: rain stopped walk, I ran, rain stop.
Doc 3: stop walking and run
Consider the query ***rain stop*** on a fictitious collection with $N = 1,000$ documents where the document frequencies of walk, rain, stop and run are respectively $50, 10, 100$ and 100. What is the similarity score for this query with documents Doc 1 and Doc 2?

Use logarithmic term weighting for query and maximum tf (term frequency) formula normalization for documents, idf weighting for the query only. Normalization is not required.

The maximum tf formula for normalization is listed as below:

$$0.25 + [0.75 \times tf_{t,d}/max(tf_{t,d})]$$

——— *End of paper* ———

Question 2

Consider following documents with the stop word list: [when, in, the, and, I]
Doc 1: when walking in the rain
Doc 2: rain stopped walk, I ran, rain stop.
Doc 3: stop walking and run
Consider the query ***rain stop*** on a fictitious collection with $N = 1,000$ documen

$idf = \log_{10} \frac{N}{df_t}$ ← 1000

	Query				Document 1				
	t_f	$w_{t,f}$	df	idf	$tf \cdot idf$	tf	$max(t_f)$	$n'd_{t,f}$	sim score
walk	0	0	50	1.3	0	1	1	1	0
rain	1	1	10	2	2	1		1	2
stop	1	1	100	1	1	0		0.25	0.25
run	0	0	100	1	0	0		0.25	0

2.25

$$n'd_{t,f} = 0.25 + 0.75 \frac{tf_{t,d}}{max\ tf_{t,d}}$$

query: { log term weighting, idf, no norm
docs: { max tf normalization, no norm

Doc 2

	t_f	$max(t_f)$	$n'd_{t,f}$	sim score
walk	1	2	0.625	0
rain	2		1	2
stop	2		1	1
run	1		0.625	0

3