1.

| | Retrieved. | Not Retrieved. | |
|---|---|---|---|
| Relevant | 4 | 6 | 10 |
| Non-Relevant. | 4 | 86 | 90. |
| | 8 | 92 | 100 |

(1) $\text{Precision} = \frac{TP}{TP+FP} = \frac{4}{8} = \frac{1}{2} = 0.5$

$\text{Recall} = \frac{TP}{TP+FN} = \frac{4}{10} = \frac{2}{5} = 0.4$

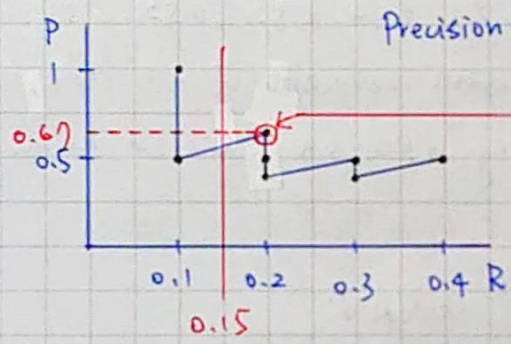
$\text{Fallout} = \frac{FP}{FP+TN} = \frac{4}{90} = \frac{2}{45} = 0.044$

$F_1\text{-score} = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2PR}{P+R} = \frac{2\times\frac{1}{2}\times\frac{2}{5}}{\frac{1}{2}+\frac{2}{5}} = \frac{4}{9} = 0.44.$

$\text{Average Precision} = \frac{1}{8}\left(\frac{1}{1}+\frac{1}{2}+\frac{2}{3}+\frac{2}{4}+\frac{2}{5}+\frac{3}{6}+\frac{3}{7}+\frac{4}{8}\right) \approx 0.5619$

(2) $[+,-,+,-,-,+,-,+] \Rightarrow$ Recall: $[0.1, 0.1, 0.2, 0.2, 0.2, 0.3, 0.3, 0.4]$

Precision: $[1, \frac{1}{2}, \frac{2}{3}, \frac{2}{4}, \frac{2}{5}, \frac{3}{6}, \frac{3}{7}, \frac{4}{8}]$




$P_{\text{inter}}(r) = \max_{r' \geq r} P(r') = \frac{2}{3} = 0.667.$

2. System A: 0 3 2 0 5 0 0 0 0 4

System B: 4 0 0 0 5 0 2 0 0 3

(1) $MAP_A = \frac{1}{4}\left(\frac{1}{2}+\frac{2}{3}+\frac{3}{5}+\frac{4}{10}\right) \doteq 0.5417$.

$MAP_B = \frac{1}{4}\left(\frac{1}{1}+\frac{2}{5}+\frac{3}{7}+\frac{4}{10}\right) \doteq 0.5571$.

(2) $R\text{-Precision} = \frac{\text{Top-}N}{N\text{- Relevant}} \Rightarrow R\text{-Precision}_A = \frac{2}{4} = \frac{1}{2} = 0.5$

$R\text{-Precision}_B = \frac{1}{4} = 0.25$

* 理論上，MAP和 R-Precision 排出來的關係會一樣，但此題的結果是相反，所以怎麼相同或不同都給對!!

(3) System A: DCG = [0, 3, 4.26, 4.26, 6.42, 6.42, 6.42, 6.42, 6.42, 7.62]

DCGI = [5, 9, 10.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89]

$\Rightarrow$ NDCG = [0, 0.33, 0.39, 0.36, 0.54, 0.54, 0.54, 0.54, 0.54, 0.64]

System B: DCG = [4, 4, 4, 4, 6.15, 6.15, 6.85, 6.87, 6.87, 7.77]

DCGI = [5, 9, 10.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89]

$\Rightarrow$ NDCG = [0.8, 0.44, 0.37, 0.34, 0.52, 0.52, 0.58, 0.58, 0.58, 0.65]

(4) We use binary (0 or 1) to represent relevance of the MAP. However, NDCG adopts the method of measuring different levels of documents with different level of relevance.

3. NNRNN RRNNR NNNRN NNNRN.

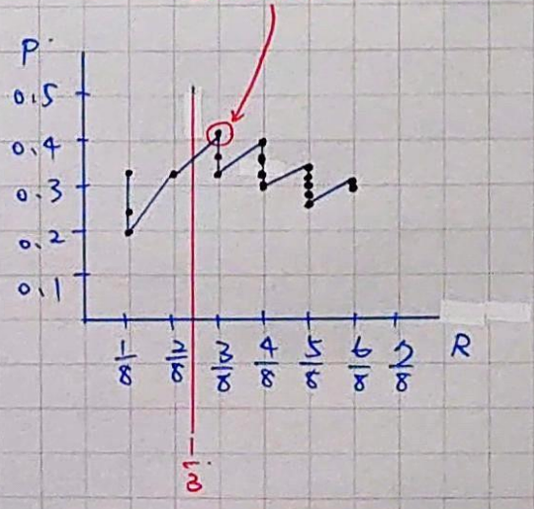
(1) Top-10 $P = \frac{4}{10} = 0.4$

(2) Top-10 $R = \frac{4}{8} = \frac{1}{2} = 0.5$ $\Rightarrow$ Top-10 F1-Score $= \frac{2PR}{P+R} = \frac{4}{9}$.

(3) $8 \times 25\% = 2 \Rightarrow P(R=25\%) = \frac{2}{6} = \frac{1}{3}$.

(4) $8 \times \frac{1}{3} = 2\frac{2}{3} \Rightarrow$ 取 3.

$\Rightarrow P_{\text{interp}}(R=\frac{1}{3}) = \frac{3}{7} \doteq 0.43$.



4. Doc 1 : recall is very very high

Doc 2 : high precision is very very very important.

(1).

| | recall | is | very | high | precision | important |
|---|---|---|---|---|---|---|
| TF(1) | 1 | 2 | 1 | 1 | 0 | 0 |
| TF(2) | 0 | 1 | 3 | 1 | 1 | 1 |
| IDF | $\log(\frac{2}{1})=0.3$ | $\log(\frac{2}{2})=0$ | $\log(\frac{2}{2})=0$ | $\log(\frac{2}{2})=0$ | $\log(\frac{2}{1})=0.3$ | $\log(\frac{2}{1})=0.3$ |
| TF-IDF(1) | 0.3 | 0 | 0 | 0 | 0 | 0 |
| TF-IDF(2) | 0 | 0 | 0 | 0 | 0.3 | 0.3 |

$\Rightarrow \vec{V_1} = [0.3, 0, 0, 0, 0, 0]$

$\vec{V_2} = [0, 0, 0, 0, 0.3, 0.3]$.

(2) Cosine Similarity = 0.

5. Doc 1 : federer has got twenty grand slam champion.

Doc 2 : nadal is the king of roland garros because he won thirteen times champion.

Query : king won champion.

(1)

| | TF(1) | TF(2) | IDF |
|---|---|---|---|
| federer – ① | 1 | 0 | 0.3 |
| has | 1 | 0 | 0.3 |
| got – ② | 1 | 0 | 0.3 |
| twenty – ③ | 1 | 0 | 0.3 |
| grand – ④ | 1 | 0 | 0.3 |
| slam – ⑤ | 1 | 0 | 0.3 |
| champion – ⑥ | 1 | 1 | 0 |
| nadal – ⑦ | 0 | 1 | 0.3 |
| is | 0 | 1 | 0.3 |
| the | 0 | 1 | 0.3 |
| king – ⑧ | 0 | 1 | 0.3 |
| of | 0 | 1 | 0.3 |
| roland – ⑨ | 0 | 1 | 0.3 |
| garros – ⑩ | 0 | 1 | 0.3 |
| because | 0 | 1 | 0.3 |
| he – ⑪ | 0 | 1 | 0.3 |
| won – ⑫ | 0 | 1 | 0.3 |
| thirteen – ⑬ | 0 | 1 | 0.3 |
| times – ⑭ | 0 | 1 | 0.3 |

(2) Jaccard Coefficient : $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

$J(Q, D_1) = \frac{|\{⑥\}|}{|\{①, ②, ③, ④, ⑤, ⑥, ⑧, ⑬\}|} = \frac{1}{8}$.

$J(Q, D_2) = \frac{|\{⑥, ⑧, ⑫\}|}{|\{⑥, ⑦, ⑧, ⑨, ⑩, ⑪, ⑫, ⑬, ⑭\}|}$

$= \frac{3}{9} = \frac{1}{3}$.

6.

(1)

| | | TF-IDF | | |
|---|---|---|---|---|
| | | $D_1$ | $D_2$ | $D_3$ |
| car | $\log \frac{10000}{18165} \doteq 0.59$ | 16.40 | 2.93 | 14.06 |
| auto | $\log \frac{10000}{6523} \doteq 1.02$ | 6.11 | 30.53 | 0 |
| insurance | $\log \frac{10000}{19241} \doteq 0.56$ | 0 | 16.83 | 16.83 |
| best | $\log \frac{10000}{25235} \doteq 0.44$ | 6.65 | 0 | 8.42 |

(2) 別以!! ① log的底數不影響排序，故小點，IDF很容易趨近1.

② 字越大量集中出現在同一文件中，TF越大，TF-IDF越容易趨近1.

(3) $\|\vec{D_1}\| = \sqrt{16.04^2 + 6.11^2 + 0^2 + 6.65^2} \doteq 18.72$

$$\Rightarrow \frac{\vec{D_1}}{\|\vec{D_1}\|} = \begin{bmatrix} 0.88 \\ 0.33 \\ 0 \\ 0.36 \end{bmatrix}$$

$\|\vec{D_2}\| = \sqrt{2.93^2 + 30.53^2 + 16.83^2 + 0^2} \doteq 34.98$

$$\Rightarrow \frac{\vec{D_2}}{\|\vec{D_2}\|} = \begin{bmatrix} 0.08 \\ 0.87 \\ 0.48 \\ 0 \end{bmatrix}$$

$\|\vec{D_3}\| = \sqrt{14.06^2 + 0^2 + 16.83^2 + 8.42^2} = 23.49$

$$\Rightarrow \frac{\vec{D_3}}{\|D_3\|} = \begin{bmatrix} 0.60 \\ 0 \\ 0.72 \\ 0.36 \end{bmatrix}$$

(4) $|IDF| = \sqrt{0.59^2 + 1.02^2 + 0.56^2 + 0.44^2} = 1.38 \Rightarrow$ Normalized IDF $= \begin{bmatrix} 0.43 \\ 0.74 \\ 0.41 \\ 0.32 \end{bmatrix}$

$\Rightarrow Q = [0.43, 0, 0.41, 0.32]$.

| | $(Q, D_1)$ | $(Q, D_2)$ | $(Q, D_3)$ | Ranking |
|---|---|---|---|---|
| Cosine Similarity | 0.73 | 0.35 | 0.98 | $D_3 > D_1 > D_2$ |
| Euclidean Distance | 18.24 | 34.76 | 22.83 | $D_1 > D_3 > D_2$ |
| Inner Product | 9.18 | 8.16 | 15.64 | $D_3 > D_1 > D_2$ |