1-1.

① precision $= \frac{4}{8} = \frac{1}{2}$ ※

② recall $= \frac{4}{10} = \frac{2}{5}$ ※

③ fall-out = False postive rate

$= \frac{FP}{TN+FP}$

$= \frac{4}{90} = \frac{2}{45}$ ※

④ $F1 = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{5}}{\frac{1}{2} + \frac{2}{5}} = \frac{4}{9}$ ※

⑤ Average precision

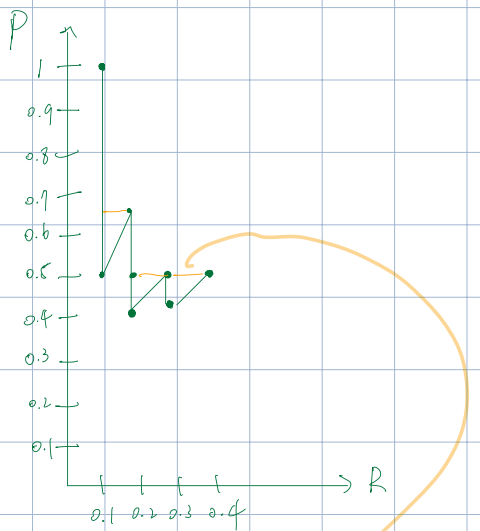
$= \frac{\frac{1}{1} + \frac{1}{2} + \frac{2}{3} + \frac{2}{4} + \frac{2}{5} + \frac{3}{6} + \frac{3}{7} + \frac{4}{8}}{8}$

$\simeq 0.5619$

1-2 ① compute top k precision & recall

P	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{2}{5}$	$\frac{3}{6}$	$\frac{3}{7}$	$\frac{4}{8}$
r	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

② plot precision & recall



③ interpolated

∴ the precision at recall level of 0.15 according to 11-interpolated

$= 0.66667$

2-1

① AP of system A : $\frac{1}{4}(\frac{1}{2}+\frac{2}{3}+\frac{3}{5}+\frac{4}{10}) \simeq 0.54$

AP of system B : $\frac{1}{4}(\frac{1}{1}+\frac{2}{5}+\frac{3}{7}+\frac{4}{10}) \simeq 0.55$

② MAP of system A $= \frac{1}{1}(0.54) = 0.54$ ✗

MAP of system B $= \frac{1}{1}(0.55) = 0.55$ ✗

2-2 ① R-precision is variation of precision at k, k become the amount of relevant $\Rightarrow \frac{|\text{retrived} \cap \text{relevant}|}{|\text{relevant}|}$

R-precision of A $= \frac{2}{4}$ ✗ R-precision of B $= \frac{1}{4}$ ✗

② No, it's not rank the system the same as MAP

2-3 Best rank : $I = \langle 5, 4, 3, 2, 0, 0, 0, 0, 0, 0 \rangle$

Best CG : $CG_I = \langle 5, 9, 12, 14, 14, 14, 14, 14, 14, 14 \rangle$

Best DCG : $DCG_I = \langle 5, 5+\frac{4}{\log_2}, 5+\frac{4}{\log_2}+\frac{3}{\log_3}, 5+\frac{4}{\log_2}+\frac{3}{\log_3}+\frac{2}{\log 4}, \cdots\cdots \rangle$

$= \langle 5, 9, 10.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89, 11.89 \rangle$

2-3 cont.

$A = \langle 0, 3, 2, 0, 5, 0, 0, 0, 0, 4 \rangle$

$CG_A = \langle 0, 3, 5, 5, 10, 10, 10, 10, 10, 14 \rangle$

$DCG_A = \langle 0, 3, 4.26, 4.26, 6.41, 6.41, 6.41, 6.41, 6.41, 7.61 \rangle$

$NDCG_A = \langle 0, 0.33, 0.39, 0.35, 0.53, 0.53, 0.53, 0.53, 0.53, 0.64 \rangle$ ※

$B = \langle 4, 0, 0, 0, 5, 0, 2, 0, 0, 3 \rangle$

$CG_B = \langle 4, 4, 4, 4, 9, 9, 11, 11, 11, 14 \rangle$

$DCG_B = \langle 4, 4, 4, 4, 6.15, 6.15, 6.86, 6.86, 6.86, 7.76 \rangle$

$NDCG_B = \langle 0.8, 0.44, 0.36, 0.33, 0.51, 0.51, 0.57, 0.57, 0.57, 0.65 \rangle$ ※

2-4

MAP is for binary judgement, while NDCG consider relevance rank ※

3.-1 #DOC = 10000, #ret DOC = 20, #Ret & RL = 6, #RL = 8

precision at 10 $= \frac{4}{4+6} = \frac{4}{10} = 0.4$ ※

On Top 10

	Ret	N-Ret	
R	4	4	8
N	6	9986	9992
	10	9990	

3.-2

Recall at 10 $= \frac{4}{4+4} = \frac{1}{2} = 0.5$

F1 at 10 $= \frac{2 \cdot \frac{4}{10} \cdot \frac{1}{2}}{\frac{1}{2} + \frac{4}{10}} = \frac{4}{9}$ ※

3.-3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P	0	0	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{2}{6}$	$\frac{3}{7}$	$\frac{3}{8}$	$\frac{3}{9}$	$\frac{4}{10}$	$\frac{4}{11}$	$\frac{4}{12}$	$\frac{4}{13}$	$\frac{5}{14}$	$\frac{5}{15}$	$\frac{5}{16}$	$\frac{5}{17}$
R	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{4}{8}$	$\frac{4}{8}$	$\frac{4}{8}$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{5}{8}$

⇓

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P	0	0	0.33	0.25	0.2	0.33	0.42	0.375	0.33	0.4	0.36	0.33	0.30	0.35	0.33	0.31	0.29
R	0	0	0.125	0.125	0.125	0.25	0.375	0.375	0.375	0.5	0.5	0.5	0.5	0.625	0.625	0.625	0.625

3-3 cont.

uninterpolated precision at 25% recall = 0.33 ✱

3-4

interpolated precision at 33% recall = 0.42 ✱

4-1

Doc 1	recall	is	very	high
tf	1	1	2	1
idf	$\log_{10}(\frac{2}{1})$	$\log_{10}(\frac{2}{2})$	$\log_{10}(\frac{2}{2})$	$\log_{10}(\frac{2}{2})$
tf-idf	0.3	0	0	0

Doc 2	precision	is	very	high	important
tf	1	1	3	1	1
idf	$\log_{10}(\frac{2}{1})$	$\log_{10}(\frac{2}{2})$	$\log_{10}(\frac{2}{2})$	$\log_{10}(\frac{2}{2})$	$\log_{10}(\frac{2}{1})$
tf-idf	0.3	0	0	0	0.3

tf-idf vector of Doc 1: (recall, is, very, high, precision, important) = (0.3, 0, 0, 0, 0, 0)

tf-idf vector of Doc 2: (recall, is, very, high, precision, important) = (0, 0, 0, 0, 0.3, 0.)

4-2

$$\text{Consine similarity} = \frac{\overrightarrow{Doc1} \cdot \overrightarrow{Doc2}}{|\overrightarrow{Doc1}| \cdot |\overrightarrow{Doc2}|} = \frac{0}{1 \cdot \sqrt{2}} = 0$$ ※

5-1

	ferderer	has	got	twenty	grand	slam	champion	nadal	is	the	king	of	roland	garrors	because	he	won	thirteen	times
Doc1:	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Doc2:	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1

※ ※

5-2

remove stop word

	ferderer	got	twenty	grand	slam	champion	nadal	king	roland	garrors	he	won	thirteen	times
Doc1:	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Doc2:	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Query:	0	0	0	0	0	1	0	1	0	0	0	1	0	0

$$\text{Jaccard}(Doc1, Query) = \frac{Doc1 \cap Query}{Doc1 \cup Query} = \frac{1}{5+1+2} = \frac{1}{8}$$ ※

$$\text{Jaccard}(Doc2, Query) = \frac{Doc2 \cap Query}{Doc2 \cup Query} = \frac{3}{9} = \frac{1}{3}$$ ※

6-1

tf-idf	Doc 1	Doc 2	Doc 3
car	$28 \cdot \log \frac{70000}{18165}$	$5 \cdot \log \frac{70000}{18165}$	$24 \cdot \log \frac{70000}{18165}$
auto	$6 \cdot \log \frac{70000}{6723}$	$30 \cdot \log \frac{70000}{6723}$	0
insurance	0	$30 \cdot \log \frac{70000}{19241}$	$30 \cdot \log \frac{70000}{19241}$
best	$15 \cdot \log \frac{70000}{25235}$	0	$19 \cdot \log \frac{70000}{25235}$

↓

tf-idf	Doc 1	Doc 2	Doc 3
car	16.40	2.92	14.06
auto	6.10	30.52	0
insurance	0	16.82	16.82
best	6.64	0	8.41
	✳	✳	✳

6-2

Yes, it can exceed 1, because both tf and idf might larger than 1 ※

6-3

$$\left\|\overrightarrow{Doc\,1}\right\|_2 = \sqrt{(16.04)^2 + (6.10)^2 + 0^2 + (6.64)^2} = 18.72$$

$$\left\|\overrightarrow{Doc\,2}\right\|_2 = \sqrt{(2.92)^2 + (30.52)^2 + (16.82)^2 + 0^2} = 34.97$$

$$\left\|\overrightarrow{Doc\,3}\right\|_2 = \sqrt{(14.06)^2 + 0^2 + (16.82)^2 + (8.41)^2} = 23.48$$

Euclidean normalised doc vector:

$\overrightarrow{Doc\,1}$: $\left(\frac{16.04}{18.72}, \frac{6.10}{18.72}, \frac{0}{18.72}, \frac{6.64}{18.72}\right) = (0.88, 0.33, 0, 0.35)$ ※

$\overrightarrow{Doc\,2}$: $\left(\frac{2.92}{34.97}, \frac{30.52}{34.97}, \frac{16.82}{34.97}, \frac{0}{34.97}\right) = (0.08, 0.87, 0.48, 0)$ ※

$\overrightarrow{Doc\,3}$: $\left(\frac{14.06}{23.48}, \frac{0}{23.48}, \frac{16.82}{23.48}, \frac{8.41}{23.48}\right) = (0.6, 0, 0.72, 0.36)$ ※

6-4

① car, auto, insurance, best

$\vec{Q} = (1, 0, 1, 1)$ $\vec{DOC1} = (1, 1, 0, 1)$

$\vec{DOC2} = (1, 1, 1, 0)$ $\vec{DOC3} = (1, 0, 1, 1)$

$\vec{Q} \cdot \vec{DOC3} > \vec{Q} \cdot \vec{DOC1} = \vec{Q} \cdot \vec{DOC2}$

∴ Rank: Doc 3 > Doc 1 = Doc 2 ※

② car, auto, insurance, best

$\vec{Q} = (0.57, 0, 0.56, 0.44)$ $\vec{DOC1} = (0.88, 0.33, 0, 0.35)$

$\vec{DOC2} = (0.08, 0.87, 0.48, 0)$ $\vec{DOC3} = (0.6, 0, 0.72, 0.36)$

$\vec{Q} \cdot \vec{DOC1} = 0.66$ $\vec{Q} \cdot \vec{DOC2} = 0.31$ $\vec{Q} \cdot \vec{DOC3} = 1.55$

∴ Rank: Doc 3 > Doc 1 > Doc 2 ※