UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

RECUPERACIÓ DE LA INFORMACIÓ

EXERCICIS DEL TEMA 4: Avaluació en recuperació de la informació

Abans de començar a resoldre els exercicis, comprova si saps respondre aquestes preguntes:

1. Explica per què són útils les col·leccions de referència i per a què s'usen.
2. Escriu les definicions de *recall* i *precision* fent servir les teves paraules.
3. Escriu la fórmula de Rocchio per recalcular una consulta a partir del *relevance feedback*. Assegura't que entens tots els paràmetres que conté.
4. Explica't a tu mateix/a com calcular les corbes de *recall*/*precision*.
5. Digues quina és la diferència entre *relevance feedback* i *pseudorelevance feedback*.
6. Per maximitzar la satisfacció de l'usuari, l'objectiu és trobar un balanç entre *recall* i *precision*. Cert o fals o discutible? Raona la teva resposta.
7. Què és un *snippet* en recuperació de la informació?

Exercici 1

Una usuària diu que, després d'haver fet una consulta al nostre sistema de cerca, ha trobat 10 documents rellevants a les posicions 2, 6, 12, 18, 20, 22, 30, 36, 40 i 50. Suposant que no hi ha més documents rellevants en la col·lecció, dibuixa una gràfica de *recall-precision* de la resposta en els 10 nivells de *recall*. Dona la taula amb els valors que has usat per dibuixar la gràfica.

Exercici 2

Tenim una col·lecció amb 100 documents identificats amb els números 1...100. Suposem que, donada una consulta, els documents rellevants són els numerats 1...20.

Dos sistemes de recuperació de la informació donen com a resultat a la consulta les respostes següents:

S1= [1, 2,21,22, 3,23,25, 4,28, 5,29,30, 6, 7,31,32,33,40,41,42, 8,43,44,
9,45,10,50,51,11,52,53,54,12,60,62,13,63,64,14,15,16,70,78,80,17,
81,82,83,85,18,90,19,91,92,20,93,94,95,96,98]

S2= [25,26, 1,27,28, 2, 3,29,30, 4,35,36, 5,37, 6,7,8,38,9,40,10,42,11,45,46,
12,48,50,51,13,60,61,64,14,70,72,15,78,79,90]

Per la consulta donada i pels dos sistemes:

1. Calcula les mesures de *recall*, *precision* i α-F-measure (amb $\alpha = 1/2$, $\alpha = 1/4$ i $\alpha = 3/4$).
2. Calcula les mesures de *coverage* i *novelty* suposant que l'usuari ja coneixia els documents de *docid* senar i no coneixia els de *docid* parell.

Exercici 3

Per la col·lecció, pregunta i sistema S1 de l'exercici anterior:

1. Dona gràfiques que mostrin el % de *recall*, precisió i precisió interpolada en funció del nombre de documents recuperats.
2. Dona la gràfica *recall-precision* pels 11 punts de *recall* estàndard (0.0, 0.1, 0.2...1.0). Nota: Recorda que la precisió interpolada en el punt de *recall* estàndard j, és el valor màxim de la precisió per qualsevol nivell de *recall* entre el j i el $(j + 1)$.
3. Calcula la precisió mitjana d'aquests 11 punts.

   Nota: La precisió mitjana és la mitjana dels valors de precisió en els punts on cada document rellevant és recuperat.

Exercici 4

Hem indexat una col·lecció de documents que contenen els termes de la taula següent; la segona columna indica el percentatge de documents en els que cada terme apareix.

Terme	% docs
computer	10%
software	10%
bugs	5%
code	2%
developer	2%
programmers	2%

Els documents estan representats internament com a vectors usant pesos *tf-idf*.

Donada la consulta $Q =$"computer software programmers", un SRI retorna els documents $D1 =$"a computer is useless without software", $D2 =$"programmers spent much of their time finding bugs in code" i $D3 =$"programmers are usually good thinkers". L'usuari considera rellevants tots tres documents $D1$, $D2$ i $D3$.

Suposa que el sistema implementa *user relevance feedback* usant la regla de Rocchio amb $\alpha = 0.8$, $\beta = 0.2$ i $\gamma = 0$. Dona, en forma de vector, la nova consulta construïda pel sistema després de la retroalimentació.