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

RECUPERACIÓ DE LA INFORMACIÓ — Control 1

Data: 16 de novembre de 2020 — **Temps: 2 hores**

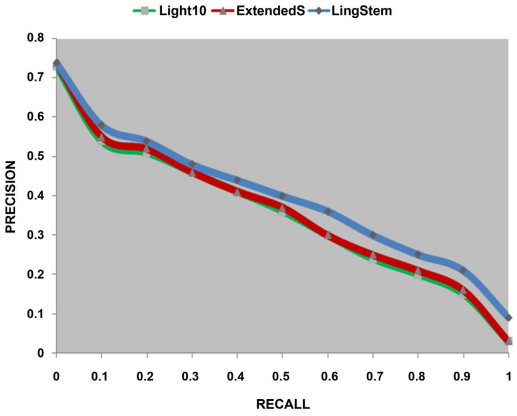
Problema 1 [2 punts]

Respon les preguntes següents, **justificant** la teva resposta:

(a) Cita almenys tres tècniques de preprocessament lingüístic que s'apliquen a un document abans de ser indexat i indica com poden afectar les mesures de *recall* i *precision*.

(b) En el rastrejador que has/heu construït, independentment del domini escollit, quines pàgines creus que no podries rastrejar o quin tipus d'informació no podries obtenir. Nota: Si no n'has construït cap, dona una resposta genèrica.

(c) La gràfica següent mostra les corbes del valor mig de *precision* en els 11 punts de *recall* estàndard sobre 15 consultes efectuades a un sistema de RI. Cada corba correspon a l'aplicació d'un *stemmer* diferent. Comenta breument el que dedueixes d'aquesta gràfica i el perquè.



(d) Com canvia (incrementa, decrementa o es manté) el valor idf_i (idf del terme t_i) en els casos següents: (1) afegint el terme t_i a un document; (2) fent que cada document dobli la seva longitud original concatenant el document amb ell mateix.

Problema 2 [3 punts]

Tenim un corpus amb un vocabulari format només per 4 termes: `a`, `b`, `c`, `d`. Les probabilitats que cada terme aparegui en un document triat a l'atzar en el corpus són: $P(a) = 0.5$, $P(b) = 0.2$, $P(c) = 0.1$ i $P(d) = 0.2$.

(a) Dona la representació dels documents $D1$ =“`a b a a d`” i $D2$ =“`a b c`” en el model vectorial usant pesos *tf-idf*. Ordena els termes alfabèticament.

(b) Ordena els documents $D1$ i $D2$ respecte la consulta $Q =$"`a b`" usant com a mesura de similitud el cosinus.

(c) Dona un exemple de dos documents no idèntics de diferents longituds amb una similitud cosinus igual a 0.

(d) Dona un exemple de dos documents no idèntics de diferents longituds amb una similitud cosinus igual a 1.

Problema 3 [3 punts]

Es vol avaluar el rendiment de dos sistemes de recuperació de la informació. Donada una consulta q, els sistemes han donat la resposta següent:

rank	1	2	3	4	5
S1	X		X		X
S2	X	X	X		

Els documents marcats amb una X són rellevants.

(a) Calcula la precisió de cada sistema quan s'han extret exactament K documents ($P@k$ o *Precision at* K) pels valors de K següents: 3, 4 i 5

(b) Creus que la mesura $P@k$ és una bona mesura per avaluar un sistema de recuperació de la informació que faci cerques a Internet? Per què?

(c) Raona què passaria si s'usés la mesura $P@k$ per una consulta que té menys de K documents rellevants com a resposta.

Problema 4 [2 punts]

Suposa que tens una xarxa de 3 pàgines enllaçades de la forma següent:

- La pàgina A té un enllaç a la pàgina C
- La pàgina B té un enllaç a la pàgina C
- La pàgina C té un enllaç a la pàgina B

Escriu la matriu de Google usant un factor d'amortiment de 0.85 i resol el sistema d'equacions resultant per calcular els valors de PageRank de les 3 pàgines de la xarxa.