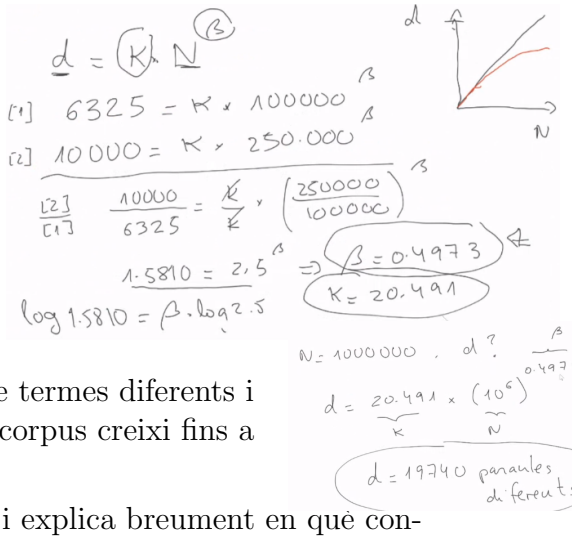
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria de Vilanova i la Geltrú

RECUPERACIÓ DE LA INFORMACIÓ — Control 1

Data: 6 de novembre de 2019 — **Temps: 1h 40m**

Problema 1 [2 punts]



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

(a) Donat un corpus que satisfà la llei de Heaps, tal que:

N	d
100000	6325
250000	10000

B = 0.4973 k = 20.491

Dóna els valors dels paràmetres estimats per descriure el nombre de termes diferents i després calcula quina serà la mida del vocabulari quan la mida del corpus creixi fins a 1000000 termes.

(b) Cita almenys tres tècniques de reducció del vocabulari i explica breument en què consisteixen. lematització, stemming, lowercase, Eliminar stopwords

(c) El càlcul de la similitud cosinus de l'angle que formen els vectors de dos documents, té en compte la normalització dels vectors (denominador de la fórmula). Per què creus que cal aquesta normalització? perque els seus valors sigin independent de la mida d'un text i aixi poder comparar textos de mida diferent.

Problema 2 [3 punts]

Donat un corpus format per 3 documents:

- Doc1: *golden sunset and golden hour*
- Doc2: *a sunset in the city*
- Doc3: *blue hour*

(a) Proposa una llista d'*stopwords* i mostra els documents després d'usar-la.

(b) Dóna la representació dels documents en el model vectorial usant pesos *tf-idf*. Ordena els termes alfabèticament.

(c) Calcula la similitud de cada document a la consulta "city hour" usant com a mesura de similitud el cosinus.

(d) Suposa que com a resultat de la consulta "city hour", l'usuari qualifica d'irrellevants els documents Doc2 i Doc3. Podries tenir en compte aquesta resposta per proporcionar un resultat millor? Com?

Problema 3 [3 punts]

Es vol avaluar el rendiment d'un sistema de recuperació de la informació. Donada una consulta q, el sistema ha donat la resposta següent:

d_{72}	d_{25}	d_{22}	d_2	d_{24}	d_{99}	d_{45}	d_{62}	d_{51}	d_4
X	X			X		X			X

Els documents marcats amb una X són els rellevants recuperats mentre que la quantitat de documents rellevants per la consulta q és 8.

A partir d'aquesta sortida, calcula les mesures següents:

(a) La precisió i el *recall* del sistema.

(b) La precisió i el *recall* per a cada posició j en la que s'extrau un document rellevant.

(c) La R-precisió, és a dir, la precisió en la posició R del rànquing de resultats per a una consulta que té R documents rellevants.

Problema 4 [2 punts]

Suposa que tenim un índex amb els termes A, B, C i D dels que coneixem la mida de les seves *postings list*: $|L_A| = 300000$, $|L_B| = 200000$, $|L_C| = 80000$ i $|L_D| = 20000$.

Recomana un pla de processament per la consulta booleana següent justificant la teva resposta:

```
(A AND C) OR (B AND D) OR (A AND D)
```