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

RECUPERACIÓ DE LA INFORMACIÓ

EXERCICIS DEL TEMA 1: Rastreig, preprocessament i estadística dels textos

---

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

1. Cita 3 sistemes de Recuperació de la Informació que usis sovint.
2. Escriu la seqüència típica de transformacions que s'apliquen a un text com a preprocessament abans de la indexació.
3. Explica la diferència entre lematització i *stemming*.
4. La llei de Zipf estableix una relació entre X i Y. Què són X i Y?
5. La llei de Heaps estableix una relació entre X i Y. Què són X i Y?

Exercici 1

Explica les tres propietats principals que ha de tenir un rastrejador. Intenta no copiar només el que està a les transparències i explica-ho amb les teves paraules.

Exercici 2

Tracta d'endevinar (sense usar cap eina) què donaria un preprocessador que eliminés paraules funcionals i fes *stemming* en el document següent:

> "Lemmatisation is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context and therefore cannot discriminate between words which have different meanings depending on part of speech. However stemmers are typically easier to implement and run faster and the reduced accuracy may not matter for some applications."
> (Font: Wikipedia)

Exercici 3

Suposeu que el nostre sistema de RI ens permet fer una consulta, que és un conjunt de paraules, i ens torna un conjunt de documents que contenen **totes** les paraules de la consulta.

Imagineu que podem configurar el sistema de quatre maneres diferents i fem quatre vegades la mateixa consulta.

- Mode 1: No eliminem *stopwords* i no fem *stemming* ni dels documents ni de la consulta. Sigui A_1 el conjunt de documents retornats.
- Mode 2: No eliminem *stopwords* però fem *stemming* dels documents i de la consulta. Sigui A_2 el conjunt de documents retornats.
- Mode 3: Eliminem *stopwords* però no fem *stemming*. Sigui A_3 el conjunt de documents retornats.
- Mode 4: Eliminem *stopwords* i fem *stemming* dels documents i de la consulta. Sigui A_4 el conjunt de documents retornats.

Quines relacions podem trobar entre A_1, A_2, A_3 i A_4? Per exemple, $A_1 = A_2$? A_2 és un subconjunt de A_4?, etc. C -> Subconjunt de...

A1 C A2
A1 C A3
A2 C A3 || A3 C A2
A1, A2, A3 C A4

Exercici 4

Tenim una col·lecció amb un total de 10^6 ocurrències de termes. Suposant que els termes es distribueixen en els texts seguint la llei de Zipf de la forma

$$f_i \cong \frac{c}{(i+10)^2}$$

doneu estimacions de (1) el nombre d'ocurrències del terme més freqüent, (2) el nombre d'ocurrències del centè terme més freqüent, i (3) el nombre de paraules que apareixen més de dues vegades. *Pista:* $\sum_{i=11}^{\infty} \frac{1}{i^2} \cong 0,095$.

Exercici 5

Ens donen una mostra aleatòria de 10 000 documents extrets d'una col·lecció que en conté 1 000 000. Comptem les paraules diferents d'aquesta mostra i en trobem 5 000. Suposant que la col·lecció satisfà la llei de Heaps amb exponent 0,5 doneu una estimació raonada del nombre de paraules diferents que espereu trobar en la col·lecció sencera.

$d = k + N^b$ $5000 = k + N^{0.5}$
$d_{col} = k * N_{con}^{0.5}$

Exercici 6

Les parelles de paraules següents es converteixen en el mateix lema un cop passades per l'*stemmer* del Porter. Quines parelles creieu que seria millor no confondre? Expliqueu per què.

1. abandon/abandonment
2. absorbency/absorbent

3. marketing/markets
4. university/universe
5. general/generate