



RECUPERACIÓ DE LA INFORMACIÓ

Control 1

Data: 14 de novembre de 2018

Temps: 1h 40m

Problema 1 [2 punts]

Responen les preguntes següents, **justificant** la vostra resposta:

- (a) L'eliminació d'*stopwords* ha de ser prèvia a l'*stemming*. Cert o fals?
- (b) Un sistema de recuperació de la informació que representa els seus documents com a vectors usant pesos *tf-idf* usa la mesura de similitud cosinus per calcular la similitud entre documents, podria recuperar un document que no contingues cap dels termes de la consulta?
- (c) Per avaluar un sistema de recuperació de la informació s'usen (entre d'altres) dues mesures: *recall* i *precision*. Per què no n'hi ha prou amb només una d'elles?
- (d) L'*Scrapy*, per defecte, està configurat per no visitar dues vegades una mateixa url. El que no evita és que es generin ítems duplicats (de diferents urls). Cert o fals?

Problema 2 [3 punts]

Calcula la similitud cosinus entre la consulta "smart tv" i el document "how to connect smart phone to smart tv". Suposa que la col·lecció consta de 10 000 000 de documents i que els termes "how" i "to" són *stopwords*.

<i>terme</i>	<i>df</i>
<i>connect</i>	3 000
<i>phone</i>	50 000
<i>smart</i>	5 000
<i>television</i>	25 000
<i>tv</i>	25 000

- (a) Què passaria si la consulta hagués estat formulada com "smart television"? Torna a calcular la similitud del document de l'enunciat amb aquesta consulta.
- (b) Proposa algun mètode per poder tractar casos com aquest.



Problema 3 [3 punts]

Considera dos sistemes de recuperació de la informació S1 i S2 que produeixen les sortides següents a 4 consultes diferents (c1...c4):

S1:	rellevants:
c1: d01 d02 d03 d04 dXX dXX dXX dXX	c1: d01 d02 d03 d04
c2: d06 dXX dXX dXX dXX	c2: d05 d06
c3: dXX d07 d09 d11 dXX dXX dXX dXX dXX	c3: d07 d08 d09 d10 d11
c4: d12 dXX dXX d14 d15 dXX dXX dXX dXX	c4: d12 d13 d14 d15
S2:	rellevants:
c1: dXX dXX dXX dXX d04	c1: d01 d02 d03 d04
c2: dXX dXX d05 d06	c2: d05 d06
c3: dXX dXX d07 d08 d09	c3: d07 d08 d09 d10 d11
c4: dXX d13 dXX d15	c4: d12 d13 d14 d15

on dXX indica que s'ha extret un document no rellevant per la consulta.

Calcula la mitjana aritmètica de les mesures de *recall* i *precision* per als dos sistemes. Digues si, en general, et sembla que calcular la mitjana d'aquestes mesures és un bon mètode per comparar dos sistemes de recuperació de la informació.

Problema 4 [2 punts]

- (a) En el context del processament d'optimització de consultes, digues quin és el criteri general per processar una consulta conjuntiva (AND).
- (b) Suposa que la teva consulta és **terme1 AND terme2 and terme3** i que disposes de les llistes de postings dels tres termes:

terme1 = 1, 11, 22
terme2 = 11, 22, 45, 72
terme3 = 29, 31, 50, 62, 90

Creus que el criteri general és òptim sempre? Raona la teva resposta.