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

RECUPERACIÓ DE LA INFORMACIÓ — Control 1

Data: 15 de novembre de 2017 — **Temps: 1h 40m**

Problema 1 [2 punts]

(a) Comenta breument quin és l'objectiu principal de cada una de les següents tasques de preprocessament: (i) eliminació d'*stopwords*, (ii) *case folding* i (iii) *stemming*.

(b) Explica breument com comprovaries que un text satisfà la Llei de Zipf.

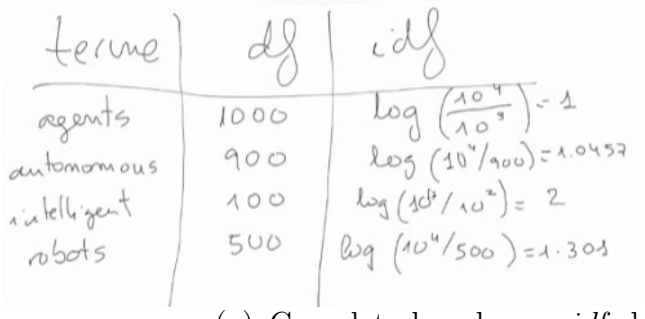
(c) Sabent que els termes `A`, `B` i `C` apareixen, respectivament, en 1000000, 500000 i 200000 documents, proposa un pla d'avaluació eficient per la consulta booleana `B and A and C`.

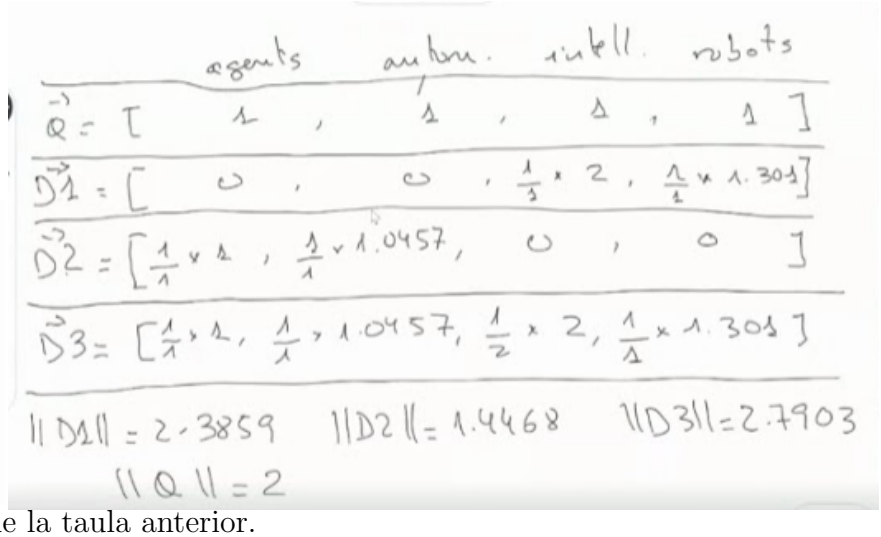
(C nad B) and A

Problema 2 [3 punts]

Hem indexat una col·lecció de 10000 documents que contenen els termes de la taula següent, on la segona columna indica el nombre de documents en els que cada terme hi apareix. Els documents estan representats internament com a vectors usant pesos *tf-idf*.

terme	*df*	*idf*
agents	1000	?
autonomous	900	?
intelligent	100	?
robots	500	?





(a) Completa la columna *idf* de la taula anterior.

(b) Donada la consulta Q ="intelligent robots and autonomous agents", un SRI retorna els documents $D1$ ="intelligent robots", $D2$ ="autonomous agents" i $D3$ ="intelligent agents and autonomous robots".

(c) Calcula la similitud de cada document amb la *query* donada usant com a mesura de similitud el cosinus.

(d) Estàs d'acord amb els resultats obtinguts? Digues alguns avantatges i alguns inconvenients sobre la mesura de similitud cosinus.

sim(D1, Q) = 0.6917
sim(D2, Q) = 0.7069
sim(D3, Q) = 0.9580

Problema 3 [2 punts]

L'Anna i el Pol necessiten escollir un sistema de recuperació de la informació que sigui adequat per les seves tasques. L'Anna treballa en una agència que es dedica a la detecció del frau en les transaccions bancàries. El Pol fa anàlisi de dades a Twitter i busca piulades que parlin malament d'un polític.

Tenim estadístiques de dos sistemes (SR1 i SR2) quan han estat utilitzats per recuperar documents en una mateixa col·lecció com a resposta a una mateixa consulta. La col·lecció contenia 100000 documents dels quals 50 eren rellevants per la consulta.

Rec1= 12/50 = 24%
Prec1= 12/15 = 80 %

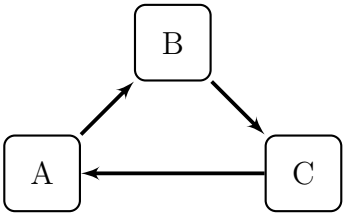
Rec2= 48/50 = 86 %
Prec2= 48/295 = 16.27%

sistema	A	R
SR1	15	12
SR2	295	48

Quin sistema recomanaries a cadascun d'ells? Raona la teva resposta.
Per la presició SR1 al Pol
Per el recall SR2 a l'Anna

Problema 4 [3 punts]

Considera la xarxa reduïda de només tres pàgines següent:



Pa = Pc
Pb = Pa
Pc = Pb
Pa + Pb + Pc = 1

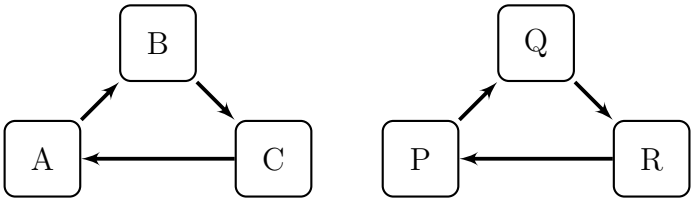
(a) Calcula el valor de PageRank de cada pàgina sense usar cap factor d'amortiment. 1/3 per cada una

(b) Una pàgina web no pot canviar el nombre de links que rep però sí que pot crear nous links que vagin a d'altres pàgines. Una pàgina web pot fer algun canvi com aquest per incrementar el seu valor de PageRank? *Pista*: Comprova què passa si C canvia el seu link cap a A i el fa anar cap a B.

Pa = 0
Pb = Pa + Pc
Pc = Pb
Pa + Pb + Pc = 1

Pb = 1/2
Pc = 1/2

(c) Suposa ara que ampliem la xarxa anterior:



$G = \lambda M + (1 - \lambda) / u * J$

Calcula el valor de PageRank de cada pàgina usant un factor d'amortiment de 0.9.