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

# RECUPERACIÓ DE LA INFORMACIÓ — Control 1

**Data: 3 de novembre de 2016** — **Temps: 1h 30min**

## Problema 1 [2 punts]

Digueu si són certes o no les afirmacions següents, justificant la vostra resposta:

(a) Un sistema de RI que recupera tots els documents de la col·lecció que té indexats, assoleix un 100% de *recall*. cert, ja que si te tots els document te totes les respostes.

(b) La Llei de Zipf ens permet estimar la freqüència de la $i$-èsima paraula més freqüent d'un corpus. cert, es just el que fa amb fi = k * r^-alpha.

(c) El component *idf* del pes *tf-idf* d'un terme també és una mesura de la raresa del terme dins el corpus. si

(d) L'*stemming* s'ha d'aplicar quan es construeix l'índex però no quan es processa la consulta. Si s'ha aplicat al index, també s'ha d'aplicar a la consulta.

## Problema 2 [3 punts]

Esteu buscant informació sobre l'augment del turisme a Barcelona en una gran col·lecció de documents. Decidiu fer la cerca usant els termes: `turisme`, `augment`, `Barcelona` i `allotjament` usant un sistema de recuperació de la informació. El sistema us recupera 3 documents. La taula següent mostra la freqüència de cada terme en cadascun dels documents:

| Termes | `allotjament` | `augment` | `barcelona` | `turisme` |
|---|---|---|---|---|
| Document 1 | 3 | 0 | 8 | 10* |
| Document 2 | 17* | 9 | 0 | 0 |
| Document 3 | 10* | 4 | 2 | 2 |
| *Query* | 1 | 1 | 1 | 1 |
| **df** | 5% | 2% | 10% | 10% |

Els valors marcats amb $*$ corresponen a la freqüència màxima dels termes en el document.

(a) Calculeu la similitud de cada document amb la *query* donada usant el sistema de pesos *tf-idf* i com a mesura de similitud el cosinus.

(b) Segons els resultats d'(a), quin document és més rellevant per aquesta *query*? Per què?

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

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

**Problema 3** [2 punts]

Suposeu que ens trobem en un escenari com el següent:

> Un hospital necessita contactar urgentment amb els pacients diabètics que van ser atesos ahir perquè tots ells han de repetir una prova que se'ls va fer malament. L'hospital usa un SRI per identificar aquests pacients. La col·lecció total d'expedients és de 10000 documents, 150 dels quals són rellevants per la consulta. El sistema recupera 250 documents, 125 dels quals són rellevants per la consulta.

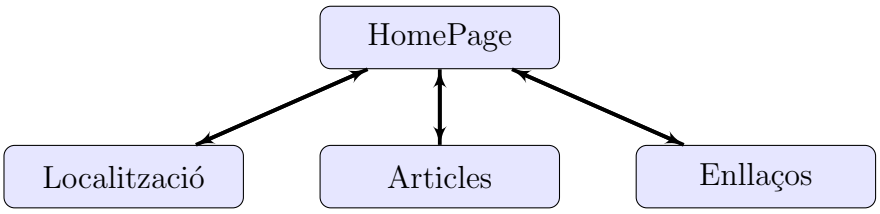
(a) Calculeu la *precision* i el *recall* del sistema. recall = 125/150 = 83% Pre = 125/250= 50%

(b) A partir dels resultats d'(a), expliqueu què representen aquestes mesures en aquest escenari en concret. Quina avaluació feu del SRI de l'hospital?

(c) Quina mesura creieu que és més important en aquest escenari? Per què?
el recall, es millor equivocarse avisant a algu que no deixarse alguna persona sense avisar.

**Problema 4** [3 punts]

Per dissenyar un lloc web, ens proposen l'estructura següent:




(a) Calculeu el valor de PageRank de cada pàgina usant un factor d'amortiment de 0.85.

(b) Creieu que el PageRank de la pàgina HomePage milloraria si el disseny fos el següent? No cal que feu els càlculs, només raoneu sobre els possibles canvis de valors de PageRank de les diferents pàgines.