Examen Final de Recuperació de la Informació
Curs 2005-2006, Quadrimestre de primavera
Temps: 2h

Feu els problemes en *fulls separats*. Es valorarà (molt) la claredat en la presentació de la resolució.

Exercici 1 (2 Punts) Sigui $K = \{k_1, \ldots, k_t\}$ un conjunt de claus i $D = \{d_1, \ldots d_\ell\}$ una col·lecció de documents. El model tf-idf permet definir el vectors de pesos $d_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,t})$ pel document $d_i \in D$. Tota la col·lecció es pot representar com la matriu

$$D = \begin{pmatrix} d_i \\ d_2 \\ \vdots \\ d_\ell \end{pmatrix} = \begin{pmatrix} w_{1,1} & \ldots & w_{1,t} \\ w_{2,1} & \ldots & w_{2,t} \\ \vdots & & \vdots \\ w_{\ell,1} & \ldots & w_{\ell,t} \end{pmatrix}$$

La similaritat de dos documents d_i, d_j es pot definir com

$$\mathsf{sim}(d_i, d_j) = \frac{\sum_{k=1}^{t} w_{i,k} \times w_{j,k}}{\sqrt{\sum_{k=1}^{t} w_{i,k}^2} \times \sqrt{\sum_{k=1}^{t} w_{j,k}^2}}$$

- (1/2 Punt) Basant-vos en la matriu D i la definició de $\mathsf{sim}(d_i, d_j)$, definiu una mesura de similaritat (o correlació) $\mathsf{sim}(k_i, k_j)$ entre les claus k_i i k_j.

  Ajut. Dos claus són similars si apareixen més o menys en els mateixos documents.

- (1/2 Punt) Demostreu $\mathsf{sim}(k_i, k_j) = \mathsf{sim}(k_j, k_i)$.

- (1/2 Punt) Calculeu $\mathsf{sim}(k_i, k_i)$.

- (1/2 Punt) Demostreu $0 \leq \mathsf{sim}(k_i, k_j) \leq 1$.

Exercici 2 (2 Punts) Per a cadascuna de les quatre tecniques següents, digueu si contribueix a millorar el Recall, la Precisió, les dues coses o cap de les dues coses.

- Stemming
- Clustering
- Relevance Feedback
- Diccionari de Sinònims

Un diccionari de sinònims per a un idioma donat indica quins termes cal considerar equivalents a l'hora de calcular la similaritat entre dos documents. Per exemple, si el diccionari indica que "espectacle" i "xou" són sinònims, es poden considerar un sol terme a l'hora de calcular distàncies.

En general, no cal que justifiqueu les respostes. Només si creieu que una tècnica no serveix per millorar ni el recall ni la precisió de les queries, intenteu explicar llavors quin és l'interés de fer-la servir.

Exercici 3 (1 Punt) Donat el text $GGACATAGATAT$ i el patró $AGACA$, simula l'execució de l'algorisme de BNDM.

Exercici 4 (1 Punt) Donat el text $aabaab$ aplica l'algorisme de SBOM per buscar els patrons aaa,$aaba$ i aba.

Exercici 5 (1 Punt)
Simuleu la cerca del patró $agra$ dins la seqüència $ararga$, on r és el caracter extés que vol dir a o g, aplicant l'algorisme de Horspool.

Exercici 6 (1.5 Punts) Construiu l'autòmata de Thomsom de l'expressió regular $ab^*(a^*|b)$ i simuleu la cerca d'aquesta expressió regular sobre el text aab.

Exercici 7 (1.5 Punts) Aplica l'algorisme de programació dinàmica per trobar el millor aliniament entre els aliniaments

$$\begin{array}{lll} a\ c\ c & \text{i} & c \\ a - c & & a. \end{array}$$