Laboratorio Sperimentale di Matematica Computazionale Lezione 3

Gianna Del Corso <delcorso@di.unipi.it>

28 Marzo 2014

1 Text Mining e Information retrieval

- Si scriva una funzione
  `function [p, r]=precrec(query_idx, cos_vector, threshold, Med_rel)`
  che dati, l'indice di una query, un vettore di coseni, un threshold e una matrice contenete la classificazione manuale di tutti i documenti rilevanti le possibili query, calcoli il valore di precision e recall, ovvero i valori

$$P = \frac{D_r}{D_t}, \qquad R = \frac{D_r}{N_t},$$

  dove D_r è il numero di documenti restituiti che sono anche rilevanti, D_t è il numero dei documenti restituiti ed N_t è il numero totale di documenti rilevanti.

- Si effettui poi una sperimentazione sulla collezione Medline prendendo i dati dalle url date in seguito.

- Si usi una sola query e si calcoli precision e recall per una opportuna sequenza di valori di soglia per il coseno. Si faccia un plot della recall vs la precision. Quando i parametri sono stati stabiliti, si calcoli e si illustri con un grafico la precision/recall su tutte e 30 le query usando il modello dello spazio vettoriale completo.

- Si calcoli la decomposizione a valori singolari sparsa della matrice termini-documenti usando il commando **svds**, e si controlli l'andamento di precision/recall del modello ottenuto approssimando la matrice con una di rango più basso (per vari valori del rango). Si scelga un valore del rango (ad esempio 100) e si ripeta l'esperimento con l'aggiunta della normalizzazione delle colonne prima di calcolare la SVD.

2 Dati

Il file MEDLINEQA.mat contiene alcuni documenti del database Medline, il file dictmed.mat contiene i termini presenti nei documenti una volta rimosse le stop words. Il file QA.mat contiene la matrice termini-documenti, le prime 30 colonne rappresentano i vettori delle query. La classificazione manuale dei documenti rilevanti si trova nel file medrel.mat. Il file può essere letto con il comando `load medrel`, se ne cancellino le ultime 2 colonne (zeri). La prima colonna rappresenta l'indice di una query, la colonna successiva un indice di un documento rilevante per quella query, cioè: `[i, j]=\mbox(medrel)(k, 1:2)` se la query i è rilevante per il documento j.