Laboratorio Sperimentale di Matematica Computazionale (Parte I) Lezione 4

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

10 Marzo 2017

1 Text Mining e Information retrieval

Per *Text mining* intendiamo metodi per estrarre informazioni utili da una collezione di testi molto grande e non strutturata. Questa operazione è conosciuta anche come information retrieval. Un'applicazione tipica è quella della ricerca di documenti all'interno di grosse collezioni scientifiche o del web. La ricerca è effettuata per mezzo di una frase di ricerca, o *query* che contiene i termini chiave per determinare i documenti rilevanti la nostra query. Il sistema deve essere in grado di fare il matching tra la query ed i documenti e presentare all'utente solo i documenti che sono rilevanti la ricerca, eventualmente ordinati secondo la rilevanza.

1.1 Descrizione del dataset

Le funzioni che implementiamo possono essere testate su un piccolo database, contenente abstract o sunti di pubblicazioni di ambito medico/biologico, estratto da MEDLINE.

I dati sono contenuti nel file `dataset_textmining` che può essere letto con il comando `load('dataset_textmining')`. Vengono caricate 4 matrici:

- Il vettore `dict_med` dei 5735 termini distinti contenuti nei documenti di Medline
- La matrice termini-documenti `A_med`, tale che `A_med(i, j)=1` se il termine `i` è contenuto nel documento `j`. I documenti sono in totale 1033.
- La matrice termini-documenti `Q_med`, contenente 30 query che utilizzano i termini in `dict_med`.
- La matrice `Med_rel`, di dimensione 30×1033 che associa, ad ognuna delle 30 query, la lista dei documenti rilevanti la query. Questa matrice è stata costruita

a mano e serve a calcolare l'indice N_r che conta il numero totale di documenti rilevanti la query, e permette di stimare la *Recall*.

Il file di testo `MEDLINEQA` contiene i documenti del database di Medline, mentre il file `common_words` la lista delle parole comuni che non sono indicizzate nella matrice `A_med`.

1.2 Modello dello spazio vettoriale: calcolo di Precision e Recall

Esercizio 1. Si scriva la funzione

```
function vc=cos_vector(A, q)
% A: matrice termini-documenti
% q: vettore di query
% vc: vettore, vc(j) memorizza il coseno dell'angolo
% tra la colonna a_j e il vettore q
```

che implementa la formula del coseno tra due vettori $\mathbf{x}, \mathbf{y}$

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

Esercizio 2. Si scriva la funzione

```
[P, R]=precision_recall(vc, query_index, epsilon, M)
% vc vettore dei coseni
% epsilon: threshold
% query_idx: indice di una query in Q_med
% M: matrice che associa ad ogni query_index la lista dei documenti rilevanti
% P: valore di Precision
% R: valore di Recall
```

che implementa il seguente schema per calcolare i valori di Precision e Recall:

1. utilizzando `vc` calcoli D_t come il numero totale dei documenti con coseno maggiore o uguale a `epsilon` (documenti restituiti);
2. calcoli, utilizzando la matrice `M`, il numero N_r di documenti rilevanti alla query;
3. calcoli D_r, cioè il numero dei documenti rilevanti restituiti come l'intersezione tra i documenti rilevanti e quelli restituiti.
4. Calcoli P e R come
$$P = \frac{D_r}{D_t}, \quad R = \frac{D_r}{N_r}.$$

Può risultare utile utilizzare le funzioni Matlab `intersect` e `find`. L'help in linea può chiarirne la sintassi e l'uso.

1. Si testi la funzione scritta con il vettore dei coseni calcolato dalla funzione precedente, con `query_idx= mod(x,30)+1` dove `x` è il proprio numero di matricola; prendendo `M=Med_rel, Q=Q_med` e facendo variare il valore di `epsilon` in un range opportuno ottenuto analizzando manualmente il vettore dei coseni.

2. Dopo aver stabilito un valore appropriato di `epsilon` per il quale D_t non sia troppo piccolo, si calcoli e si faccia un grafico della media precision/recall su tutte e 30 le query usando il modello dello spazio vettoriale completo.

Esercizio 3. Si scriva una funzione

```
function [vc]=LSI(A, q, k)
% A: matrice terminixdocumenti
% q: vettore query
% k: intero
% vc: vettore dei coseni
```

che implementi il seguente algoritmo:

1. normalizzi per colonna la matrice A;

2. calcoli la SVD troncata al k-esimo valore singolare della matrice termini-documenti normalizzata[1];

3. restituisca `vc`, il coseno dell'angolo tra la proiezione del vettore `q` sul nuovo spazio dei documenti e la proiezione delle colonne sullo spazio dei documenti. Si svolga questa operazione in modo da lavorare su vettori di lunghezza k.

Si scriva uno script che dopo aver invocato funzione `LSI(A, q, k)` con valori di k=25, 50, 75, 100, calcoli la media aritmetica di precision e recall su tutte e 30 le query del database Medline. Si confrontino questi valori con quelli di precision e recall ottenuti al punto 2 dell'esercizio 2 utilizzando lo spazio vettoriale completo.

[1] Si usi il commando **svds** con parametro k che opera su matrici sparse