Sistemi informativi (corso progredito)
a.a. 2014/2015

Laboratorio n. 6

Latent Semantic Analysis

Massimo Melucci

Obiettivi

- Comprensione di Latent Semantic Analysis (LSA).
- Un'implementazione di LSA.
- Una funzione di reperimento che integra LSA.

Come funziona

Dopo aver implementato l'algoritmo e calcolato LSA per la collezione, ci sono diversi modi per integrare LSA, ad esempio:

1. si produce il primo *ranking* (vedi laboratorio n. 2) con o senza RF, a scelta (vedi laboratorio n. 4)
2. si fissa un numero $N \geq 1$ e si usano i primi N documenti reperiti
3. si produce una matrice di occorrenza di dimensione $k \times N$ dove k è il numero di descrittori trovati negli N documenti
4. si calcola LSA
5. si proietta il vettore dell'interrogazione sul sottospazio del primo autovettore (quello con l'autovalore più grande)
6. si riordinano i primi N documenti
7. gli altri documenti sono accodati ai primi N riordinati

Base di partenza

- Pacchetti di decomposizione di matrici disponibili per R, Matlab, Python, ecc., ad esempio ci sono i seguenti (non provati, da usare a vostro rischio e pericolo):
  - pacchetto `lsa` per R (`http://cran.r-project.org/web/packages/lsa/lsa.pdf`)
  - pacchetto `gensim` per Python (`http://radimrehurek.com/gensim/`)
  - pacchetto Sequential Latent Semantic Indexing per Matlab (`http://www.mathworks.com/matlabcentral/fileexchange/22795-sequential-latent-semantic-indexing`)
- I dati della lezione precedente.

Procedimento

Il procedimento è in un due passi, uno per ogni consegna:

- Alla prima consegna si deve produrre un documento di testo che illustra la propria funzione di reperimento integrata con LSA. Il nome del documento deve essere nel seguente formato: `lab-6-gruppo-`*n*`.txt`, dove *n* è il numero del gruppo; il formato può anche essere Microsoft Word (si usi l'estensione `.doc` o `.docx`) o Adobe PDF (si usi l'estensione `.pdf`).
- continua...

Procedimento

- ...
- Alla seconda consegna, si devono produrre i seguenti risultati:
  - La versione definitiva del testo fornito alla prima consegna.
  - I risultati dell'esecuzione della funzione di reperimento nel seguente formato testuale:

    Id.Int. `Q0` Id.Doc. Rango Punteggio EtichettaRun

    dove EtichettaRun è l'etichetta che identifica la *run* nel formato `G`*n*`R`*m*`LSA` dove *n* è il numero del proprio gruppo, *m* è il numero della run da 0 a 9 e `LSA` sta per Latent Semantic Analysis; ad esempio `G17R3LSA` è la *run* n. 3 del gruppo n. 17 con LSA
- continua...

Procedimento

- ...
- Ad esempio:

```
1   Q0   1234     1   1.234   G17R3LSA
1   Q0   345      2   1.056   G17R3LSA
1   Q0   2909     3   1.056   G17R3LSA
...
1   Q0   12     114   0.056   G17R3LSA
1   Q0   879    115   0.034   G17R3LSA
1   Q0   3204   116   0.023   G17R3LSA
2   Q0   12       1   3.467   G17R3LSA
2   Q0   879      2   3.123   G17R3LSA
...
```

- **Attenzione**: non si aggiunga l'intestazione appena descritta.
- Si metta la *run* in un file di testo omonimo, ad esempio, `G17R3LSA.txt`
- Si carichi un archivio compresso con all'interno tutti i file.