Sistemi Informativi
Laboratorio 5

Catalin Copil Mattia de Stefani Giulio Lovisotto

May 7, 2015

1 Descrizione

Utilizzeremo la funzione di ranking BM25 con relevance feedback esplicito (lab 5), e applicheremo LSA. Sceglieremo i primi N documenti reperiti (proveremo vari valori di N per trovare quello che garantisce il miglior risultato). Costuiremo la matrice di occorrenza ridotta X di dimensione $(k \times N)$ dove k e' il numero di descrittori trovati negli N documenti. Useremo il metodo `linalg.svd` della libreria `numpy` per trovare la fattorizzazione :

$$X = U\Sigma V^T.$$

Dopodiche', considereremo m dimensioni (proveremo vari valori di m a partire da 1 per trovare quello che garantisce il miglior risultato), e proietteremo l'interrogazione $\vec{q}$ sullo spazio a dimensione ridotta usando la formula:

$$\vec{q}_m = \Sigma^{-1} U_m^T \vec{q},$$

faremo lo stesso con i documenti computanto la matrice ridotta $X^* = U_m \Sigma_m V_m^T$. Poi computeremo la cosine similarity tra i documenti nello spazio ridotto e la query ridotta, e li riordineremo per cosine similarity descrescente. Il ranking degli altri documenti oltre agli N non verra' modificato, ed essi verranno accodati.