Sistemi Informativi
Laboratorio 5

Catalin Copil Mattia de Stefani Giulio Lovisotto

May 13, 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). Useremo il metodo `linalg.svd` della libreria `numpy` per trovare la fattorizzazione della matrice termini-documenti X:

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

Dopodiche', considereremo m dimensioni, e proietteremo l'interrogazione $\vec{q}$ sullo spazio a dimensione ridotta usando la formula:

$$\vec{q}_m = \Sigma_m^{-1} U_m^T \vec{q}.$$

Poi computeremo le cosine similarity tra le rappresentazioni degli N documenti nello spazio ridotto (colonne di V_m^T) e la query ridotta $\vec{q}_m$. Ordineremo questi N documenti per cosine similarity decrescente, mentre l'ordine di tutti gli altri documenti non verra' alterato, e questi verranno accodati.

2 Risultati

Abbiamo utilizzato i seguenti parametri, $m = 2$, $N = 10$. Nel report finale proveremo diverse configurazioni per trovare il miglior risultato. Figura 1 riporta i valori di map ottenuti.

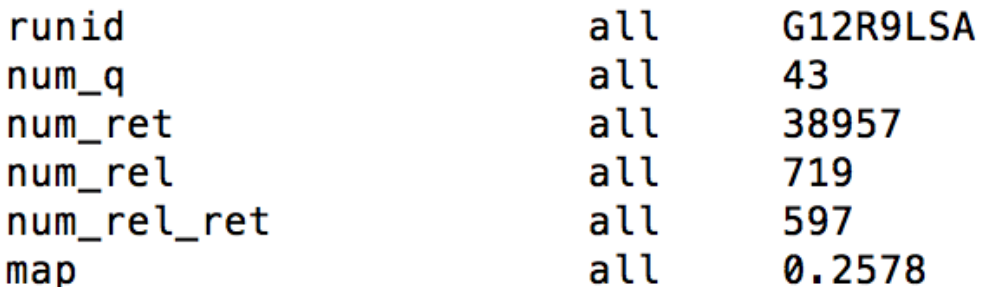

Figure 1: Map utilizzando LSA.