Laboratorio 3

Catalin Copil Mattia de Stefani Giulio Lovisotto

April 16, 2015

Descrizione

Abbiamo scelto di usare l'algoritmo di ranking BM25 . Tale algoritmo funziona nel modo seguente:

$$\sum_{i \in Q} \log\Big(\frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}\Big) \cdot \frac{(k_1 + 1) f_i}{k + f_i} \cdot \frac{(k_2 + 1) q f_i}{k_2 + q f_i}$$

dove:

- i sono i termini della query Q,
- r_i e' il numero di documenti rilevanti che contiene il termine i,
- R e' il numero di documenti rilevanti per la query,
- n_i e' il numero di documenti che contiene il termine i nella collezione,
- N e' il numero totale di documenti nella collezione,
- k_1, k_2 sono parametri,
- f_i e' la frequenza del termine i nel documento,
- qf_i e' la frequenza del termine i nella query,
- K e' definito nel modo seguente:

$$K = k_1((1 - b) + b \cdot \frac{dl}{avdl})$$

dove b e' un parametro, dl e' la lunghezza del documento, $avdl$ e' la lunghezza media di un documento nella collezione. Questo termine serve a normalizzare il componente di frequenza rispetto alla lunghezza del documento (per non favorire i documenti troppo lunghi).

I termini R e r_i sono informazioni note a priori, tipicamente sono settati a zero in quanto non si hanno informazioni sulla rilevanza. Nel nostro caso li ignoriamo in un primo momento lasciandoli a zero.

Implementazione

Descriviamo ora le strutture dati necessarie al reperimento che vengono calcolate durante l'indicizzazione.

- matrice delle frequenze n_docs x n_words
- un vettore che contiene le lunghezze dei documenti

La funzione di reperimento scorrera' la lista di documenti (le righe della matrice), e per ogni documento scorrera' sui termini della query Q (Document-at-a-time retrieval). Poi i documenti con punteggio maggiore di zero verranno ordinati (dal maggiore al minore).

I prossimi passi saranno quelli di tenere in considerazione le informazioni di rilevanza contenute nel file *qrels.txt* per ottimizzare il reperimento (abbiamo gia' i documenti rilevanti per ogni query). Useremo l'utility *trec_eval* per scegliere la configurazione che produce il miglior risultato.