INFORMATION RETRIEVAL: HOMEWORK 1

Mario Avdullaj, matricola 1184284
DEI, Università degli Studi di Padova

23 Gennaio 2019

1 Introduzione

Presentiamo in questa relazione una sintesi dei risultati ottenuti sull'analisi della collezione TREC7, collezione composta da oltre 500mila documenti, 50 topic e i relativi relevance assessment. Durante la sperimentazione, sono state effettuate quattro run utilizzando diverse indicizzazioni e modelli per il reperimento. Infine, i risultati ottenuti da ciascuna run sono stati sottoposti al test statistico one-way ANOVA.

2 Sistema IR

Per condurre gli esperimenti sopracitati ci siamo avvalsi del sistema di reperimento Terrier, alla versione 4.4. La fase di indicizzazione è avvenuta settando opportunamente il file di settaggio *terrier.properties* con gli opportuni parametri, quali la *termpipelines*, la quale permette l'inclusione o meno (nel processo di indicizzazione) delle stop list e stemming. Le fasi di retrieval ed evaluation sono state poi automatizzate con degli script, variando da linea di comando la ranking function e la termpipelines, nonché la lista dei topic e le qrels. Infine ci siamo avvalsi di uno script di parsing per la raccolta dei risultati delle varie run. Precisando, i dati rilevanti che sono stati analizzati sono la *average precision (AP)*, *precision at 10 (P@10)* e *Recall precision (Rprec)*, dati utilizzati quindi da script matlab per il test ANOVA ed il plottaggio dei grafici.

3 Dati sperimentali

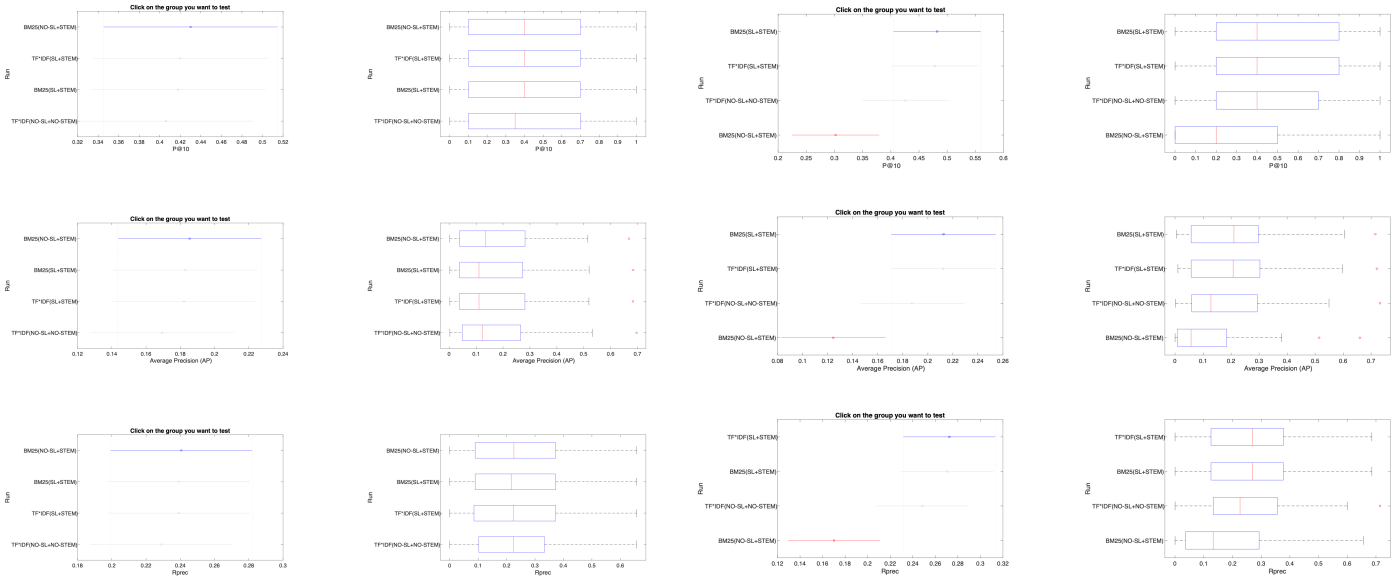
Veniamo dunque ai risultati sperimentali delle nostre run. Inizialmente state effettuate quattro run considerando come topic solo i campi TITLE. In seguito ai risultati ottenuti, abbiamo provato ad utilizzare query più dettagliate, aggiungendo quindi anche il campo DESC. In questo modo volevamo constatare come variassero le prestazioni dei modelli utilizzati, in quanto il campo DESC contiene descrizioni più esaustive dei documenti che vogliamo reperire, contenendo anche parole ripetute, o di poca rilevanza. I risultati ottenuti sono i seguenti:

Topic:TITLE	**AP**	**Rprec**	**P@10**	**num_rel_ret**
BM25 (Stoplist + Stemmer)	0.1828	0.2391	0.4180	2277
BM25 (No stoplist, Stemmer)	0.1854	0.2406	0.4300	2287
TF*IDF (Stoplist + Stemmer)	0.1821	0.2391	0.4200	2264
TF*IDF (No stoplist, No stemmer)	0.1693	0.2290	0.4060	2064

Topic:TITLE+DESC	AP	Rprec	P@10	num_rel_ret
BM25 (Stoplist + Stemmer)	0.2125	0.2705	0.4820	2586
BM25 (No stoplist, Stemmer)	0.1245	0.1701	0.3020	1403
TF*IDF (Stoplist + Stemmer)	0.2123	0.2725	0.4780	2577
TF*IDF (No stoplist, No stemmer)	0.1876	0.2485	0.4260	2315

Analizzando i risultati delle prime quattro run - in particolare AP ed il numero di documenti rilevanti recuperati - è interessante osservare come le prestazioni di ciascun sistema non si discostano molto dalla media. Utilizzando query più dettagliate, aggiungendo il campo di descrizione, otteniamo invece prestazioni generalmente migliori, migliorando del 15/20% i vari coefficienti di valutazione tenuti in considerazione. L'unico sistema che invece ha ottenuto netti peggioramenti di prestazione risulta il BM25 senza utilizzo di stoplist e porter stemmer. Si presume dunque che le parole utilizzate nei campi DESC abbiano avuto un coefficiente IDF basso, diminuendo lo score dei documenti della collezione e quindi il numero finale di documenti recuperati.

Dopo aver sottoposto i risultati ottenuti dalla fase di evaluation al test statistico one-way ANOVA e HSV, sono stati ottenuti i seguenti risultati:



(a) Runs con TITLE

(b) Runs con TITLE, DESC

Dal test statistico possiamo affermare come, per il caso a), possiamo accettare l'ipotesi nulla di ANOVA, ossia che non vi sono differenze statistiche significative fra le diverse run per i diversi parametri considerati, potendo perciò considerare le medie di ciascuna di esse uguale per il top-group considerato. Diverso è il caso b), ove il test ANOVA evidenzia come la run con il BM25 (senza stoplist) abbia prestazioni nettamente inferiori, scartando dunque l'ipotesi nulla del test statistico. Concludendo, abbiamo constatato che generalmente fare uso di query corte, con frasi brevi ma significative, come il TITLE, ci assicura prestazioni tutto sommato stabili e modeste (considerando gli indici e modelli utilizzati nell'esperienza). Inoltre è possibile evidenziare come il modello TF*IDF abbia ottenuto prestazioni simili per quanto riguarda le run sia con stoplist e stemmer sia senza queste ultime. Ciò invece non è avvenuto per il modello BM25, lasciandoci presumere che termini molto frequenti e di bassa rilevanza influenzino molto la pool finale di documenti recuperati per questo modello.