Homework 1

Nicolò Levorato 1156744

1 Software utilizzato

Per l'homework sono stati utilizzati i seguenti strumenti:

- Terrier 4.4 per indicizzare e interrogare la collezione.
- Trec_eval per valutare i risultati.
- Script Matlab per condurre i vari test.

2 Lavoro svolto

Per interrogare ed indicizzare la collezione sono è stato utilizzato il software Terrier. Sono state eseguite quattro run con le seguenti specifiche:

- Stoplist, Porter stemmer, BM25.
- Stoplist, Porter stemmer, TF*IDF.
- No stoplist, Porter Stemmer, BM25.
- No stoplist, No stemmer, TF*IDF.

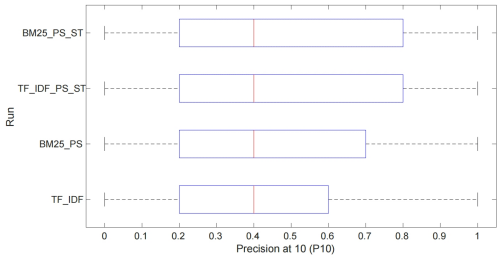
Le varie istruzione sul metodo di indicizzazione ed interrogazione sono state fornite al sistema attraverso i vari file properties. In particolare sono state fatte due interrogazioni differenti per ogni run. Una in cui si è considerato solo "title" nel campo dei TREC topics, un'altra in cui si è considerato sia "title" sia "desc". In entrambi i casi si è preferito ignorare i valori con IDF troppo basso per avere risultati migliori.

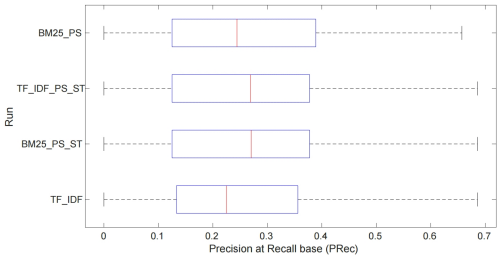
In seguito sono stati analizzati i risultati ottenuti Trec_eval. I file output di Trec_eval sono stati utilizzati per ricavare la matrice data.mat grazie allo script Matlab collect.m.

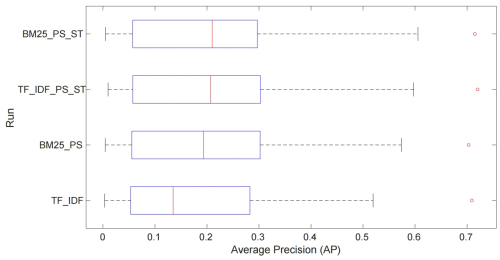
La matrice data.mat è stata utilizzata come input per i test Anova 1-way effettuati grazie agli script:

- anovamap.m per "Mean Avarage Precision"
- anovarprec.m per "Precision At Recall Base"
- anovap10.m. per "Precisio At 10"

I grafici riportati in seguito fanno riferimento solamente alle interrogazioni in cui si considerano entrambi "title" e "desc".







Run	Map	RPrec	P_10
BM25_ps_st	0.2126	0.2705	0.4840
TFIDF_ps_sl	0.2120	0.2725	0.4800
BM25_ps	0.2108	0.2740	0.4740
TFIDF	0.1875	0.2460	0.4300
P value	0.8471	0.7886	0.7836

3 Conclusioni

I test Anova 1-way verificano la H0 hypotesis poiché tutti i p value sono maggiori di 0,5, quindi le run sono statisticamente equivalenti. Grazie al test HSD di Tuckey si può vedere inoltre che tutte le run appartengono al top group.

4 Confronto tra "TITLE" e"TITLE,DESC"

La tabella mostra i risultati dell'interrogazione che tiene in considerazione solo "title":

Run	Map	RPrec	P_10
BM25_ps_st	0.1828	0.2391	0.4180
TFIDF_ps_sl	0.1821	0.2391	0.4200
BM25_ps	0.1857	0.2409	0.4300
TFIDF	0.1693	0.2290	0.4060

Confrontando questi valori con quelli ottenuti in precedenza si vede che le performance del sistema peggiorano. Questo è dovuto al fatto che il sistema, avendo meno "informazioni" da confrontare con la query, fa più fatica a recuperare documenti rilevanti. L'unico motivo per preferire utilizzare solo "title" è se si ha bisogno di diminuire i tempi di retrieval a scapito della precisione.