

Specifiche software

Per eseguire l'homework ho usato il seguente software:

- *Terrier 4.4* per l'indicizzazione e l'interrogazione;
- *trec_eval 9.0* per la valutazione;
- Script *Matlab* per l'analisi.

Descrizione del lavoro

Per l'homework ho eseguito quattro run usando *Terrier*, indicizzando e interrogando la collezione opportunamente seguendo le specifiche date, ovvero:

- **BM25_SL_PS**: Stoplist, Porter stemmer, BM25
- **TFIDF_SL_PS**: Stoplist, Porter stemmer, TF*IDF
- **BM25_PS**: No stoplist, Porter Stemmer, BM25
- **TFIDF**: No stoplist, No stemmer, TF*IDF

Per le query ho considerato entrambi i campi "*title*" e "*desc*" dei *TREC Topic* ignorando valori con IDF troppo basso a favore di un miglioramento considerevole delle prestazioni nella run *BM25_PS*. (Vedi tabelle in basso)

In seguito ho valutato i risultati delle run eseguite con *Terrier* usando *trec_eval*.

Per l'analisi del sistema ho usato i seguenti script e file *Matlab*.

- *textEvalToData.m* converte in due file *.mat* i valori quali MAP, precision at recall base e precision at 10 dai risultati della valutazione di *trec_eval*.
- *evalData.mat* e *evalAllData.mat* sono file dati risultanti da *textEvalToData.m* e contengono rispettivamente i valori per ogni singolo topic e i valori calcolati per l'intera run.
- *displayTable.m* stampa una tabella per confrontare i valori tra le quattro diverse run. Utilizza il file dati *evalAllData.mat*.
- *ir_anova_map.m*, *ir_anova_rprec.m* e *ir_anova_p_10.m* servono per condurre il test *ANOVA 1-way* sulle varie misure. Utilizza il file dati *evalData.mat*.

Preciso che ho eseguito il test *ANOVA 1-way* per verificare l'equivalenza dei sistemi solo per le run che ignoravano termini con valori bassi di IDF. Boxplot e tabella con i dati del test *ANOVA 1-way* sottostante si riferiscono quindi solamente a queste.

Risultati sperimentali & Test ANOVA

Run name	MAP	Rprec	P_10
BM25_SL_PS	0.2126	0.2705	0.4840
TFIDF_SL_PS	0.2120	0.2725	0.4800
BM25_PS	0.2108	0.2740	0.4740
TFIDF	0.1875	0.2460	0.4300

Tabella 1: Risultati ottenuti ignorando termini con basso IDF

Run name	MAP	Rprec	P_10
BM25_SL_PS	0.2125	0.2705	0.4820
TFIDF_SL_PS	0.2123	0.2725	0.4780
BM25_PS	0.1245	0.1701	0.3020
TFIDF	0.1876	0.2485	0.4260

Tabella 2: Risultati ottenuti non ignorando termini con basso IDF

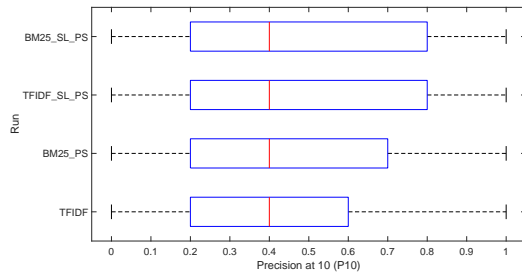
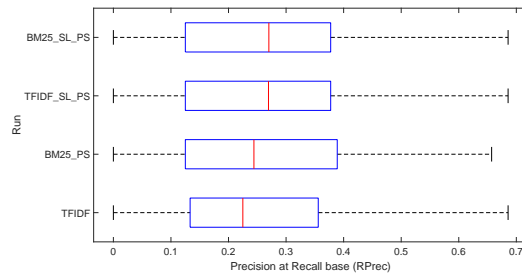
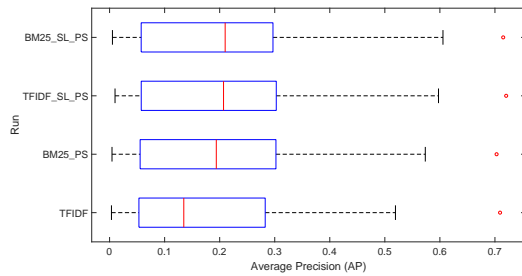


Figura 1: Boxplots

ANOVA	MAP	Rprec	P_10
F	0.2698	0.3508	0.3578
p value	0.8471	0.7886	0.7836

Figura 2: ANOVA 1-way

Considerazioni sui risultati: ANOVA 1-way

A seguito del test *ANOVA 1-way* possiamo dire che le run sono statisticamente equivalenti. Si nota infatti che tutti i *p value* sono maggiori di 0.05, verificando quindi la *null hypothesis H0*. Eseguendo il test HSD di Tukey (script matlab per la versione interattiva presente nella repository) si può provare che tutte le run appartengono al *top-group*.

Considerazioni aggiuntive sui risultati: ignorare termini con basso IDF?

Ignorare o meno i termini con valore basso di IDF generalmente non cambia in modo rilevante la performance del sistema per le run di questo homework. L'unica eccezione si ha per la run *BM25_PS* dove le prestazioni per la *MAP* ad esempio cambiano da 0.21 a 0.12. Il risultato è coerente con quanto studiato nella teoria: in una run per cui in fase indicizzazione non ho filtrato *stop words* sarà più probabile recuperare documenti non rilevanti che contengono quelle parole solo perchè di uso comune.