

Homework 1

Studente: Giovanni Barbieri

Matricola: 1177495

Repository: https://github.com/linparkkin/IR_HW1

Introduzione

Lo svolgimento del seguente homework consiste nell'analisi della collezione sperimentale *TREC7* composta da circa 52800 documenti, 50 topic e un pool con due gradi di rilevanza: R, NR. L'analisi consiste nell'utilizzo di uno strumento di Information Retrieval, in questo caso *terrier* v4.4, per lo sviluppo di 4 diverse run:

- stoplist, porter stemmer, BM25;
- no stoplist, porter stemmer, BM25;
- stoplist, porter stemmer, TF*IDF;
- no stoplist, no porter stemmer, TF*IDF;

Valutare poi le run calcolando MAP, Rprec e Precision at 10 utilizzando un tool di valutazione, in questo caso *trec_eval*, già disponibile all'interno di *terrier*.

Condurre alla fine il test statistico ANOVA 1-way per determinare i sistemi appartenenti al "top group" sulla base delle diverse misure.

Valutazione

Prima di tutto è stato utilizzato *terrier* per effettuare l'indicizzazione dei documenti presenti nella collezione. Per fare ciò è stato modificato il file *terrier.properties* del tool, cambiando di volta in volta in volta i parametri in base alla run che si era interessati a sviluppare. Va specificato che nella configurazione utilizzata in questo homework sono stati presi in considerazione anche i termini con un basso indice di frequenza *idf*, impostando *ignore.low.idf.terms = false*. Una volta ottenuti i 4 diversi indici è stato possibile valutare le run utilizzando *trec_eval* dal terminale, lanciando il comando `sh trec_eval.sh -q -m map -m Rprec -m P.10 qrels.trec7.txt`. Non si è tenuto conto della descrizione dei topic settando *TrecQueryTags.skip = DESC, NARR*. La seguente tabella riassume i risultati ottenuti.

	map	Rprec	P@10
sp-BM25: stoplist, porter stemmer, BM25	0.1828	0.2391	0.4180
np-BM25: no stoplist, porter stemmer, BM25	0.1854	0.2406	0.4300
sp-TFIDF: stoplist, porter stemmer, TF*IDF	0.1821	0.2391	0.4200
nn-TFIDF: no stoplist, no porter stemmer, TF*IDF	0.1693	0.2290	0.4060

Table 1: Risultati della valutazione.

Come si può vedere la run che ha ottenuto i risultati migliori è np-BM25, con lo score migliore in tutti e tre i criteri di valutazione. Va tuttavia evidenziato come i valori delle quattro diverse run siano molto vicini fra loro, a esclusione di nn-TFIDF che rappresenta il caso peggiore sia nella Mean Average Precision che nella Recall Precision e Precision at 10.

ANOVA 1-way

Lo step finale consiste nella realizzazione del test statistico ANOVA 1-way. A tal fine sono stati sviluppati con Matlab gli scripts *anova.m*, che contiene l'implementazione del test, *parser.m*, realizzato per effettuare il parsing delle misurazioni fornite da *trec_eval*, e *run-plots.m*, contenente il codice per stampare i grafici delle diverse run; tutti i file sono disponibili all'interno della repository. Per un miglior confronto dei risultati del test ANOVA, è stato implementato il test Tukey HSD, che fornisce un'intuitiva rappresentazione grafica. Figure 1 e Figure 2 illustrano i risultati ottenuti dai test.

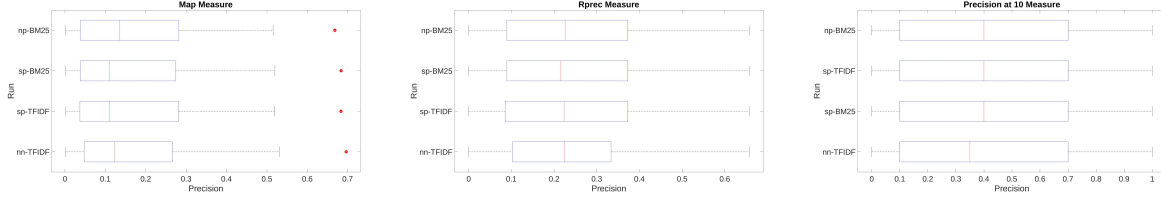


Figure 1: Boxplot per le 3 diverse misure di precisione.

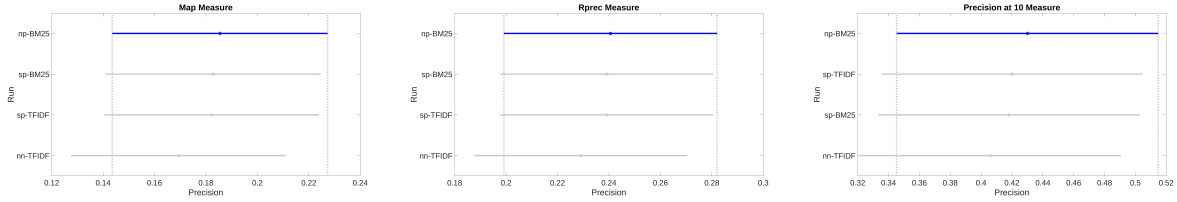


Figure 2: Tukey HSD test per le 3 diverse misure di precisione.

I grafici riassumono quanto già parzialmente evidenziato nella sezione riguardante la valutazione, ovvero che np-BM25 ha una performance migliore rispetto agli altri casi analizzati. I boxplots sottolineano come np-BM25, sp-TFIDF e sp-BM25 abbiano una media molto simile, soprattutto nel caso della Precision at 10, dove i relativi valori della mediana e degli altri quartili sono identici, rispettivamente $q_{1/4} = 0.1$, $q_{2/4} = 0.4$ e $q_{3/4} = 0.7$. Il mantenimento delle stopwords e il mancato utilizzo del Porter stemmer, influenzato probabilmente anche dalla scelta di questa analisi di prendere in considerazione i termini con un basso indice di frequenza *idf*, ha invece diminuito la performance nel modello di pesatura TF*IDF, infatti in questo caso $q_{2/4} = 0.35$. Tale andamento si replica anche per la Mean Average Precision, nonostante gli outliers dovuti al topic 15, e la Recall Precision, dove nn-TFIDF ha un miglior primo e secondo quartile rispetto a sp-BM25 e sp-TFIDF, ma un terzo quartile nettamente inferiore.

Il test di Tukey conferma np-BM25 come la soluzione con la performance migliore rispetto alle altre 3, dovuto alla scelta di non tenere in considerazione il tag *DESC* dei topic nel processo di valutazione e al non aver rimosso le stopwords dai documenti della collezione. Tukey evidenzia inoltre come np-BM25, sp-TFIDF e sp-BM25 abbiano una media molto simile, mentre nn-TFIDF è quello che si discosta più di tutti.

Altri interessanti plot sono disponibili all'interno della repository.