Information Retrieval

HOMEWORK 1

Pignotti Matteo
1179803

Introduzione

In questo homework è stato utilizzato il software Terrier (versione 4.4.2). Si è considerata, in fase di indicizzazione, la collezione sperimentale TREC7 composta da circa 528000 documenti, 50 topic e un pool con 2 gradi di rilevanza.
Utilizzando questo software di reperimento sono state eseguite quattro run secondo le specifiche dettate dall'homework, specificando che, i vari file terrier.properties, sono stati modificati solo nel parametro "termpipelines" mentre cambio di modello (TF*IDF, BM25) è avvenuto tramite riga di comando.
In riferimento alla valutazione, invece, è stato utilizzato il Trec_eval integrato da Terrier, tenendo in considerazione anche i termini con basso IDF. Ottenuti poi i vari indici di valutazione, essi sono stati salvati in un file txt.
Successivi test ANOVA sono stati sviluppati tramite codice matlab visto a lezione.

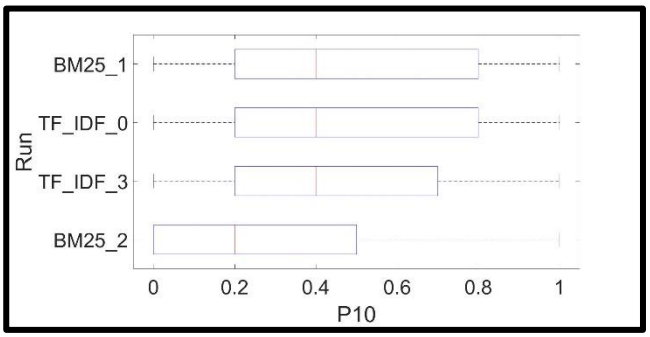
Test Statistico & ANOVA

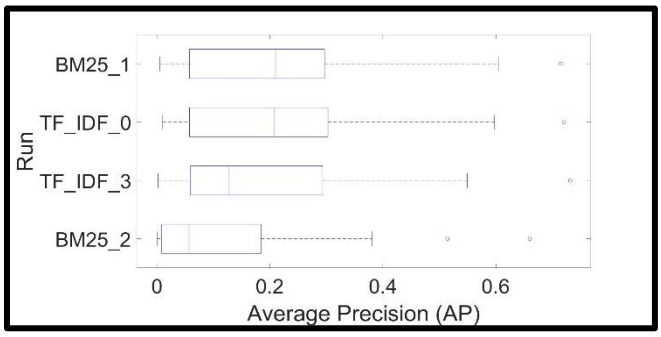
Dopo aver usato Trec_eval e parsato i vari risultati sono stati estrapolati i dati necessari e condotte le analisi opportune, inerenti a l'AP per topic, Rprec e P10, per ogni singola run.
Vorrei sottolineare di aver istanziato due blocchi di run diverse; la prima non tiene conto del campo "desc", quindi considera solo il titolo del topic, la seconda invece valuta e indicizza anche le relative descrizioni. Ho potuto constatare che, come si può notare dalle due tabelle seguenti, solo il BM25 senza stoplist aveva una MAP, Rprec e P_10 inferiore nel secondo blocco (run con desc). Molto probabilmente questo deriva dal fatto che i risultati vengono falsati dalla presenza di termini ad alta frequenza ma a basso contenuto informativo, che verrebbero invece esclusi inserendo la stop list. Ciò porta il modello ad assegnare un alto score a documenti non effettivamente rilevanti che verranno poi "scartati" da Trec_eval dopo il confronto che avviene con i qrels.

Senza desc	Num_rel	Map	Rprec	P_10	Recall_10	nDCG_10
BM25 Stoplist e PS	2277	0.1828	0.2391	0.4180	0.0895	0.4432
BM25 No Stoplist e PS	2287	0.1854	0.2406	0.4300	0.0901	0.4509
TDIDF Stoplist e PS	2264	0.1821	0.2391	0.4200	0.0909	0.4444
TDIDF No Stoplist e No PS	2064	0.1693	0.2290	0.4060	0.0843	0.4251

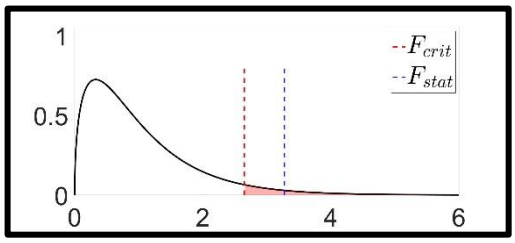
Con parametro desc	Num_rel	Map	Rprec	P_10	Recall_10	nDCG_10
BM25 Stoplist e PS	2586	0.2125	0.2705	0.4820	0.1000	0.5167
BM25 No Stoplist e PS	1403	0.1245	0.1701	0.3020	0.0699	0.3213
TFIDF Stoplist e PS	2577	0.2123	0.2725	0.4780	0.1008	0.5144
TFIDF No Stoplist e No PS	2315	0.1876	0.2485	0.4260	0.0873	0.4570

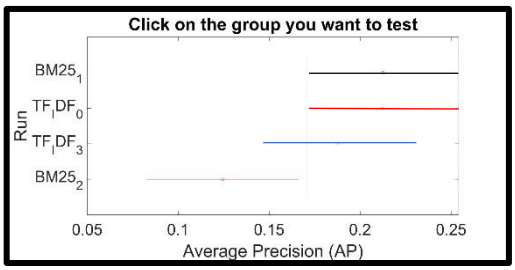
Le future analisi verranno fatte su run dove è stato considerato il campo description dato che nel primo caso non abbiamo delle differenze sostanziali tra sistemi.





Da questi due Plot possiamo notare come inizialmente le varie *AP distribution* siano sbilanciate verso valori abbastanza bassi. Ciò indica che molto probabilmente, all'aumentare dei documenti rilevati, lo score ad essi assegnato è in disaccordo con i giudizi di rilevanza dati dai qrels. Questa considerazione la si può trarre dal boxPlot della Precision_10 che si aggira attorno ad un valore di 0.4 per tutte le run ad eccezione del BM25 senza stoplist.





Basandomi sull'ANOVA analysis, invece, si può notare come almeno in un caso c'è una differenza sostanziale tra le varie medie dei vari gruppi di analisi. Come visto a lezione, considerando la null Hyphotesis come l'ugualianza tra le medie delle rispettive distribuzioni, e analizzando il Test Statistico, possiamo vedere come F_{stat} sia maggiore del F_{crit}. Questo mi fa intuire che è possibile quindi rifiutare l'ipotesi H_0.
Grazie al Test di Tukey e quindi un confronto tra coppie di run, riusciamo ad identificare inoltre il sistema che si discosta maggiormente, in media, dagli altri tre: il BM25 senza stoplist (la motivazione sarà più chiara al termine dell'analisi che segue).

Concludendo si può notare come nel caso del BM25, una run senza stoplist, vada a influire molto negativamente sulla precision, rispetto ad una basata sul modello TF*IDF. Questo avviene molto probabilmente dal fatto che la decisione sulla rispettiva rilevanza di un documento o meno, dipende dalla frequenza di quel determinato termine all'interno dell'intera collezione dei documenti rilevanti. L'assenza della stoplist va ad influire negativamente sul pull di termini, cosa che si riscontrerà poi sul calcolo delle le varie probabilità. Quest'ultime in conclusione mi serviranno nel considerare o meno un documento rilevante, osservandone all'interno, la loro presenza. Termini come articoli e proposizioni, andranno a influire in maniera più drastica nel reperimento nel BM25 rispetto al TF*IDF (metodo di pesatura nato per ovviare già inizialmente a questo problema e quindi sviluppato per sottovalutare tutte quelle words molto frequenti sia nel documento che nella collezione, che in molti casi sono proprio appartenenti alla stoplist). Quello che mi rassicura è la scarsa variazione che ho della precisione sulle due run della TF*IDF senza e con stoplist.
Per questioni di spazio ho messo solo i grafici a mio parere più significativi del test Anova. Altri plot si possono trovare all'interno della cartella relativa al Homework.