Relazione Homework 1

Corso: Reperimento dell'Informazione A.A. 2018-2019
Candidato: Rudy Berton
Matricola: 1157241
Repository: https://github.com/rudyberton92/IR_Homework1

Software e librerie

Per lo svolgimento del progetto é stato utilizzato il software Terrier v.4.4 nella fase di indicizzazione e reperimento sulla collezione sperimentale TREC7 fornita. La valutazione invece é stata svolta con i software Trec_Eval v.9.0 e Matlab v.9.5, il primo per ottenere le misure, mentre il secondo per svolgere il calcolo del test statistico Anova 1-way e il test di Tukey HSD.

Scelte implementative

É stato scelto di porre a confronto due simulazioni di reperimento dell'informazione: quella classica di un utente comune che sfrutta i motori di ricerca per ottenere delle informazioni e quella invece impiegata da un utente maggiormente esperto e consapevole di cosa stia cercando e pertanto capace di effettuare una ricerca piú accurata. Questa distinzione ha richiesto unicamente il cambio dei parametri riguardanti i topic presenti nel file *terrier.properties*: per simulare un utente comune che utilizza un numero limitato di termini nella barra di ricerca del browser é stato scelto di interrogare la collezione di documenti unicamente con il campo *title* presente nei topic; diversamente sono stati usati i campi *title* e *description* (*desc*) per simulare l'interrogazione da parte di un utente con un'esigenza informativa piú delineata, capace quindi di fornire al sistema anche dettagli del contesto ricercato. Queste due scelte implementative porteranno ad ottenere risultati ampiamente diversi in fase di valutazione, come si vedrá in seguito.
Altri parametri presenti nel file *terrier.properties* invece sono stati mantenuti costanti per entrambe le simulazioni. É stato scelto di non ignorare i valori bassi di IDF (*ignore.low.idf.terms=false*, giá di default nel sistema Terrier) dal momento che non ho ritenuto sufficientemente ampio il corpus di documenti in possesso, decidendo quindi di tenere in considerazione anche i termini che potrebbero presentarsi piú frequentemente nella collezione. Rimangono inoltre invariate le proprietá riguardanti l'analisi lessicale, quali la codifica in UTF-8 dei documenti e la tokenizzazione in lingua Inglese, unica lingua presente nella collezione.
Per entrambe le simulazioni sono state eseguite le stesse quattro run, ognuna con caratteristiche diverse per quanto riguarda l'indicizzazione e il reperimento utilizzati da Terrier (Figure 1). Per la fase di rimozione delle stopwords viene utilizzata la stop-list offerta di default da Terrier, mentre per il passo di stemming l'algoritmo utilizzato é quello di Porter. I modelli impiegati per il reperimento sono due: TF-IDF e BM25, dove quest'ultimo viene eseguito coi valori di default usati dal sistema (k1=1.2 e b=0.75).

	Analisi lessicale	Stopwords	Stemming	Modello
BM25b0.75 - 1	Si	Si	Si	BM25
TF-IDF - 2	Si	Si	Si	TF-IDF
BM25b0.75 - 3	Si	No	Si	BM25
TF-IDF - 4	Si	No	No	TF-IDF

Figure 1: Run con relative caratteristiche

Valutazione

Nella fase di valutazione oltre alle misure richieste (*MAP*, *Rprec* e *Precision at document cut-off* con valore k=10), si é valutata la capacitá del sistema nel reperimento di documenti (*num_ret*), il numero totale di documenti considerati rilevanti (*num_rel*) e il numero totale di documenti rilevanti che sono stati reperiti (*num_rel_ret*). Figure 2 riporta i valori calcolati su tutti i topic.
Focalizzandosi unicamente sui valori dei sistemi relativi alla simulazione di un utente comune é possibile notare come, a fronte dello stesso modello TF-IDF usato, l'impiego di stemmer e stop-list (`TF-IDF-2`) produca un aumento del numero di documenti rilevanti ritornati rispetto al non utilizzo di tali tecniche (`TF-IDF-4`). Stessa nota va fatta anche per i medesimi sistemi nella simulazione di un utente esperto. Completamente diverso é la situazione per quanto riguarda il confronto tra i sistemi con il modello probabilistico BM25 (`BM25b0.75-1` e `BM25b0.75-3`). Se nel caso di utente comune la differenza tra i due risulta relativamente minima (con valori maggiori per `BM25b0.75-3`), la differenza nel caso dell'utente esperto risulta molto marcata: l'utilizzo di molti termini nei topic (titolo e descrizione) senza la rimozione delle stopwords porta ad una diminuzione quasi del 50% del numero di documenti rilevanti ritornati. Questo viene confermato anche dal confronto tra i due `BM25b0.75-3`, relativi alle due simulazioni, registrando una differenza

	Utente comune				Utente esperto				Differenza: utente esperto - utente comune			
	BM25b0.75-1	TF-IDF-2	BM25b0.75-3	TF-IDF-4	BM25b0.75-1	TF-IDF-2	BM25b0.75-3	TF-IDF-4	BM25b0.75-1	TF-IDF-2	BM25b0.75-3	TF-IDF-4
num_ret	47396	47396	47396	45826	50000	50000	50000	50000	+ 2604	+ 2604	+ 2604	+ 4174
num_rel	4674	4674	4674	4674	4674	4674	4674	4674	0	0	0	0
num_rel_ret	2277	2264	2287	2064	2586	2577	1403	2315	+ 309	+ 313	- 884	+ 251
MAP	0.1828	0.1821	0.1854	0.1693	0.2125	0.2123	0.1245	0.1876	+ 0.0297	+ 0.0302	- 0.0609	+ 0.0183
Rprec	0.2391	0.2391	0.2406	0.2290	0.2705	0.2725	0.1701	0.2485	+ 0.0314	+ 0.0334	- 0.0705	+ 0.0195
P@10	0.4180	0.4200	0.4300	0.4060	0.4820	0.4780	0.3020	0.4260	+ 0.0640	+ 0.0580	- 0.1280	+ 0.0200

Figure 2: Misure per i singoli sistemi in entrambe le simulazioni e loro differenza

in negativo significativa per tutte le misure.
Nel caso studio di un utente comune, analizzando i dati delle altre misure, é possibile notare come i valori dei quattro sistemi risultino abbastanza vicini tra loro; il sistema che presenta le performance migliori si rivela essere quello con il modello probabilistico BM25 ma senza la rimozione delle stopwords (`BM25b0.75-3`). Ció non vale invece nel caso di un utente esperto: in tale simulazione é possibile vedere come proprio il sistema migliore del caso precedente presenti valori significativamente lontani rispetto agli altri, presentando le peggiori prestazioni in tutte le misure. Questa volta le prestazioni migliori si ottengono nei sistemi che implementano sia la rimozione delle stopwords, sia il Porter Stemmer: valori leggermente superiori per il modello BM25 (`BM25b0.75-1`) nelle misure *MAP* e *P@10*, mentre un valore maggiore per la *Rprec* nel modello TF-IDF (`TF-IDF2`). Ció sta ad indicare come l'utilizzo di un maggior numero di termini per ogni topic in fase di reperimento, con un'opportuna indicizzazione, sia piú efficace rispetto all'uso di solo pochi termini.
I valori ottenuti per le diverse misure trovano conferma anche dal test statistico Anova 1-way e dal test di Tukey HSD in entrambe le simulazioni. Seguono le osservazioni riguardanti tali test eseguiti sui valori dell'*Average Precision* (*AP*) di ogni singolo topic; i test calcolati sulle misure *Rprec* e *P@10* presentano risultati concordi con quelli esposti qui e sono reperibili all'interno del repository.
Con il calcolo dell'Anova si assume come ipotesi (H0) che le medie dei quattro sistemi siano uguali tra loro in entrambe le simulazioni svolte. Dai risultati ottenuti, nel caso di un utente comune, non c'é evidenza per rifiutare tale ipotesi: dal grafico boxplot (Figure 3.a) infatti é possibile vedere come le mediane siano tutte vicine tra loro e non ci sia una significativa variazione tra i sistemi, tutti infatti rientrino nel top-group secondo il grafico ottenuto col test di Tukey (Figure 3.b). Nella simulazione di un utente esperto, il boxplot mostra come un sistema presenti una mediana completamente spostata rispetto alle altre (Figure 4.a); le medie dell'*AP* (ossia *MAP*) pertanto non sono tutte uguali arrivando a rifiutare l'ipotesi di partenza (H0). Con il test di Tukey, attraverso i confronti a coppie tra i sistemi é possibile identificare quale sia quello significativamente differente agli altri (Figure 4.b): come ci si poteva aspettare giá dai valori visti precedentemente é per l'appunto il sistema `BM25b0.75-3` che non rientra nel top-group.

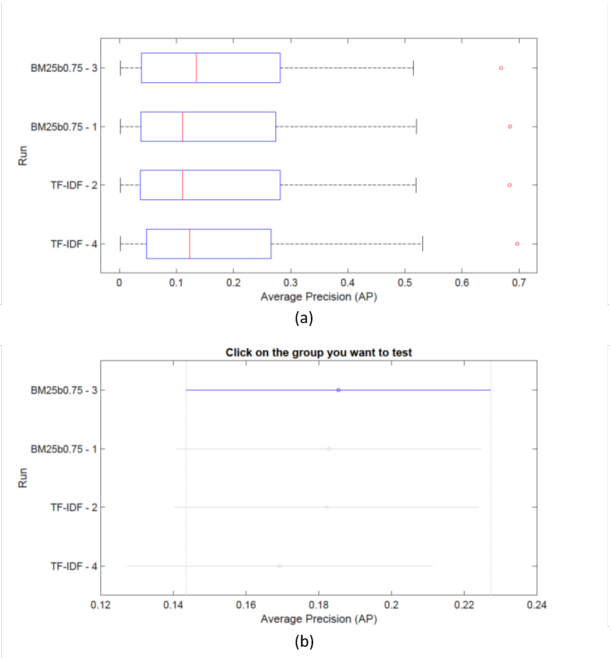

Figure 3: Caso utente comune - boxplot e test di Tukey

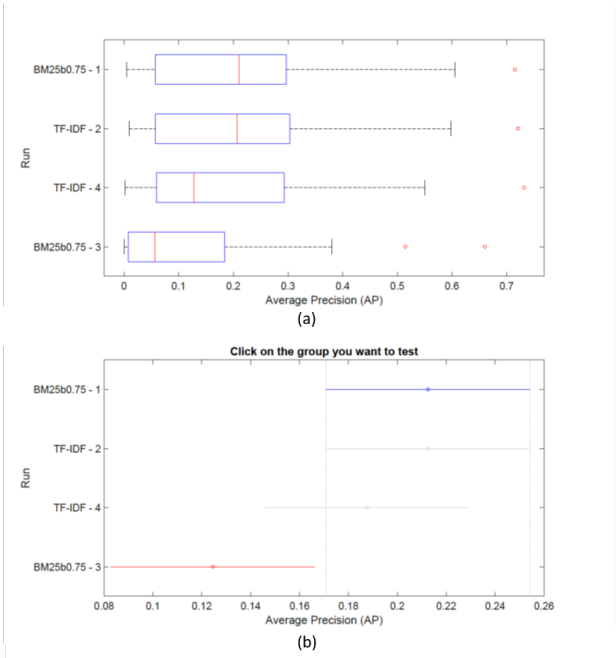

Figure 4: Caso utente esperto - boxplot e test di Tukey

Conclusioni

Se all'interno di ogni singola simulazione i dati raccolti e valutati sono concordi ai rispettivi risultati ottenuti dai test, questo non risulta vero per quanto riguarda il confronto diretto tra i due casi di reperimento analizzati. L'introduzione di un maggior numero di termini nei diversi topic (campi *title* e *desc*) ha portato il miglior sistema a diventare il peggiore del proprio gruppo, sotto le stesse condizioni. Il sistema `BM25b0.75-3` pertanto, a differenza degli altri, non risulta consistente in variazione al numero di termini utilizzati per il reperimento. Con gli altri sistemi invece si ha un'evidente miglioramento delle performance, fornendo come query anche un contesto di ció che viene cercato.
Come avevo potuto presumere ad inizio lavoro, un reperimento svolto con un maggior numero di termini permette la raccolta di piú documenti rilevanti che vanno a soddisfare l'esigenza informativa dell'utente, in questo caso di un utente esperto. Bisogna prestare attenzione peró al processo iniziale di indicizzazione utilizzato per ottenere migliori prestazioni dal sistema: con i modelli utilizzati e il corpus a disposizione la fase di indicizzazione richiede sia la rimozione delle stopwords, sia l'uso dell'algoritmo di Porter per lo stemming.