Information Retrieval - Homework1

Tommaso Sgarbanti

Università di Padova
tommaso.sgarbanti@studenti.unipd.it

January 12, 2019

I. Introduzione

Per questo progetto è stata utilizzata la collezione sperimentale TREC7 composta da circa 528000 documenti, 50 topic e un pool con 2 gradi di rilevanza: R, NR (Relevant, Non-Relevant). Utilizzando Terrier come sistema di Reperimento dell'Informazione sono state eseguite le seguenti run:

- BM25 con stoplist e PorterStemmer
- BM25 senza stoplist e con PorterStemmer
- TF*IDF senza stoplist e con PorterStemmer
- TF*IDF senza stopList e senza stemmer

Le run sono state poi valutate sulla base delle seguenti misure: Mean Average Precision, R-Precision (Precisione a un livello di cut-off pari alla recall base) e della Precision@10 (Precisione a un livello di cut-off 10). Infine sono stati utilizzati i test statistici ANOVA 1-way e Tukey test per osservare eventuali differenze significative tra le run, i risultati saranno presentati e discussi nei paragrafi successivi. Per riferirci alle run verrà utilizzata la seguente simbologia: +PS=è stato utilizzato PorterStemmer, +SL=è stata utilizzata la stoplist, -SL=non è stata utilizzata la stoplist, -S=non è stato utilizzato lo stemmer.

Link alla repository:
https://github.com/sgarbanti/IR2018-19-HW1

i. Strumenti utilizzati

Per l'indicizzazione e il reperimento dei documenti è stato utilizzato Terrier v4.4., tra le impostazioni del file properties è importante specificare che, per quanto concerne la fase di reperimento, sono stati considerati i tag "title" e "desc" e ignorato invece il tag "narr". Per la valutazione delle run è stato invece utilizzato il tool trec_eval.9.0, strumento standard utilizzato nella comunità TREC. Infine i test statistici sono stati condotti utilizzando Matlab R2018b.

i.1 Valutazione delle run

Indicizzati, per ciascuna diversa configurazione, i documenti della collezione ed effettuato il reperimento di essi sulla base delle 50 query, sono state valutate le quattro run utilizzando diverse misure di valutazione. Nella tabella 1 seguente sono mostrati i risultati sulla base della Mean Average Precision, R-Precision e della Precision@10.

	MAP	R-Prec	Prec@10
BM25 +SL +PS	0.2126	0.484	0.2705
BM25 -SL +PS	0.2108	0.474	0.2740
TF*IDF +SL +PS	0.2120	0.480	0.2725
TF*IDF -SL -S	0.1875	0.430	0.2460

Table 1: *Risultati Valutazione*

L'analisi di questi dati permette alcune considerazioni:

- Modello BM25:
  - Utilizzare una stoplist permette un reperimento che posiziona un maggior numero di documenti rilevanti in posizione alte di rank, rispetto che ad effettuare l'indicizzazione dei documenti senza l'ausilio di essa. Va però notato come la precisione a cut-off 10 sia invece più alta nella run senza stoplist, ciò significa che nelle prime 10 posizioni più alte di rank, nella

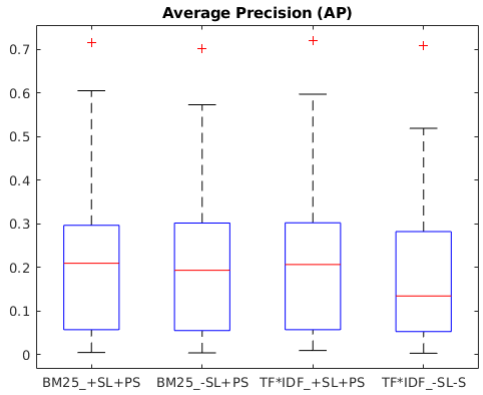

Figure 1: *BoxPlot Average Precision*

lista di documenti reperiti, vi è un numero maggiore di documenti rilevanti rispetto che alla run che utilizza la stoplist.

- Modello TF*IDF: Il non utilizzo ne di una stoplist ne di uno stemmer per l'indicizzazione dei documenti porta, in fase di reperimento, a una diminuzione generale delle performance.

Tra i due modelli non è possibile fare grandi osservazioni, nelle due run in cui sarebbe possibile confrontarli, quelle che utilizzano entrambe sia la soptlist che PorterStemmer, hanno ottenuto valutazioni molto simili. Per avere una visione grafica delle valutazioni sono stati prodotti dei boxplot, uno per ciascuna misura, in cui si può osservare come in base a topic diversi i modelli si siano comportati più o meno bene, alla figura 1 si può osservare il boxplot riguardante l'Average Precision. Il boxplot ci mostra come in base al topic variano, anche in maniera significativa, le valutazioni. Si può però osservare, come anche visto nella tabella precedente, che non appaiono differenze di performance significative tra le diverse run, se non forse per la run TF*IDF senza stoplist e senza stemmer che presenta una mediana più bassa rispetto alle altre.

ii. ANOVA 1-way

Per accertarsi che non vi siano realmente sostanziali differenze tra le run è stato condotto il test statistico Anova 1-way per ciscuna misura di valutazione, in questo modo è si è testata l'ipotesi nulla che la media delle misure sulle diverse run sia uguale contro l'ipotesi alternativa che, in una o più run, differiscano. Nella tabella 2 è possibile osservarne i risultati.

	SS	DF	MS	F-static	p-value
MAP	0.0223	3	0.0074	0.2698	0.8471
R-Prec	0.0264	3	0.0088	0.3508	0.7886
Prec@10	0.0938	3	0.0313	0.3578	0.7836

Table 2: *ANOVA 1-way test*

Il valore più interessante da osservare è il p-value che per tutte e tre le misure si trova con valori vicino a 1, essendo un p-value alto non vi è alcuna evidenza empirica contro l'ipotesi nulla. Il test statistico condotto non ha dunque mostrato alcun indizio per il quale si possa pensare che vi sia qualche modello che si differenzi dagli altri in in maniera significativa.

iii. Tukey HSD Test

Il Tukey Honestly Significant Difference (HSD) test compara due gruppi alla volta, secondo tutte le possibili combinazioni, creando un intervallo di confidenza per ciascun confronto, ad un certo livello di significatività α, al cui interno si trova il valore vero della differenza tra le medie dei due gruppi con una confidenza pari a $(1-\alpha)\%$. In questo caso sarebbe servito per individuare quali fossero le run che differivano dalle altre, visto che dal test ANOVA 1-way si può capire se ve ne sono ma non quali sono. Come detto però il test ANOVA 1-way non ha mostrato una tale situazione, il Tukey test è stato ugualmente eseguito confermando quanto già discusso precedentemente. Nella tabella 3 sono riportati i risultati del test per quanto riguarda la misura Mean Average Precision (MAP). In tutti i confronti vi sono p-value alti vicini a 1, inolte per tutte le coppie gli intervalli di confidenza, calcolati con un liv-

1-RUN	2-RUN	L-LIMIT	DIFF	U-LIMIT	P-VALUE
BM25 +SL +PS	TF*IDF +SL +PS	-0.0847	0.0005	0.0858	1.0000
BM25 +SL +PS	TF*IDF -SL +PS	-0.0835	0.0018	0.0870	0.9999
BM25 +SL +PS	TF*IDF -SL -S	-0.0602	0.0251	0.1103	0.8740
TF*IDF +SL +PS	BM25 -SL +PS	-0.0840	0.0012	0.0865	1.0000
TF*IDF +SL +PS	TF*IDF -SL -S	-0.0607	0.0246	0.1098	0.8808
BM25 -SL +PS	TF*IDF -SL -S	-0.0619	0.0233	0.1086	0.8958

Table 3: *Tukey HSD Test*

ello di significatività $\alpha = 0.05$, comprendono il valore 0 e i limiti dell'intervallo, "L-LIMIT" e "U-LIMIT" sono anche vicini a esso.

iv. Conclusioni

Quattro diverse run sono state ottenute utilizzando la collezione TREC7, le quattro run in questione sono: BM25 con stoplist e PorterStemmer, BM25 senza stoplist e con PorterStemmer, TF*IDF senza stoplist e con PorterStemmer, TF*IDF senza stoplist e senza stemmer. Le run sono state inizialmente valutate secondo le misure di Mean Average Precision, R-Precision e Precision@10, poi sono stati condotti dei test statistici quali ANOVA 1-way e Tukey test. Dai risultati ottenuti si può evincere che nessuna run si differenzia in maniera significativa dalle altre, infatti anche guardando le misure di valutazione non si osservano grandi differenze. L'evidenza maggiore c'è con il modello TF*IDF dove, eliminando l'utilizzo della stoplist e di PorterStemmer, si ha avuto un calo di prestazioni riscontrato sulla base di tutte e tre le misure di valutazione utilizzate.