

# [IR2018-19-HW1] 1183535 Simone Nigro

## 1. Lavoro svolto

Per svolgere l'homework è stato utilizzato l'ambiente *macOS Mojave* (versione 10.14.2).

Si è inoltre fatto uso di *terrier-core* (versione 4.4), *trec\_eval* (versione 9.0.4) e del programma *PyCharm* (versione 2018.3.1).

Verranno presentate, in questa sezione, le diverse fasi in cui si è suddiviso il lavoro (reperibile insieme ai risultati ottenuti su <https://github.com/nsimone93/Information-Retrieval>):

- inizializzazione del software *terrier* attraverso il comando `./trec_setup.sh` dando in input la posizione della cartella *TIPSTER* contenente il nostro data set;
- nel file *terrier.properties* sono state inserite le proprietà *trec.topics* (indicante la posizione del file *topics.351-400\_trec7.txt* contenente il topic relativo ai documenti), *trec.qrels* (con la posizione del file *qrels.trec7.txt* contenente i giudizi di rilevanza), *ignore.low.idf.terms* (impostato al valore *true* per ignorare valori di *idf* bassi), *TrecQueryTags.process* (impostato a *title,desc* in modo da considerare sia il titolo che la descrizione del topic nelle query) e *TrecQueryTags.skip* (impostato a *narr* per evitare di considerare la narrazione e quindi un appesantimento del carico di lavoro nell'esecuzione delle query);
- indicizzazione tramite il comando `./trec_terrier.sh -i` (nel caso in cui fosse necessaria un'indicizzazione che non facesse uso del *Porter stemmer* o della rimozione delle *stopword*, prima di questa fase veniva modificato il file *terrier.properties* in cui si impostava correttamente la proprietà *termpipelines*);
- dopo aver creato l'indice, tramite il comando `./trec_terrier.sh -printstats`, è stato possibile verificare le statistiche risultanti e si è constatato che il numero di documenti indicizzati ammontava a 528155;
- esecuzione delle run tramite il comando `./trec_terrier.sh -r -Dtrec.model="Tipo_modello"` utilizzando i sistemi descritti in *Tabella 1*;

Nome sistema	Modello BM25	Modello TF_IDF	Stop list	Porter stemmer
BM25	✓		✓	✓
TF_IDF		✓	✓	✓
BM25_porter	✓			✓
TF_IDF_not		✓		

Tabella 1. Sistemi utilizzati

- ottenuti i file *.res* da ogni run eseguita, è stata eseguita una fase di valutazione tramite il software *trec\_eval* usando il comando `./trec_eval -q -m all_trec` fornendo in input il file *qrels.trec7.txt* con i giudizi di rilevanza e il *.res* da valutare; si sono ottenute così tutte le misure che verranno poi analizzate;
- creazione di uno script in *Python 3.7*, utilizzando il programma *PyCharm*, per inserire in una struttura le misure di valutazione da analizzare. Sono stati creati così *Average\_Precision.txt*, *P\_10.txt* e *Rprec.txt* contenenti per ogni file una matrice di 50 righe (Topic) e 4 colonne (Run) indicanti la relativa misura di valutazione in esame;
- vengono effettuati l'ANOVA 1-way ed il *Tukey HSD test*, utilizzando i tre file appena creati, riportando i risultati nei file *Anova\_ "Misura\_considerata".txt* e *tukey\_HSD\_ "Misura\_considerata".txt*;
- viene infine generato il plot della MAP per ogni sistema analizzato e vengono rappresentate le misure *P(10)* e *Rprec* creando un plot per ogni run.

Per la creazione dello script si è fatto uso della libreria *os* per le chiamate di sistema, *statsmodel* per l'ANOVA 1-way e per il *Tukey HSD test*, *matplotlib* per la creazione dei plot ed infine le librerie accessorie *scipy* e *numpy*.

## 2. Analisi dei risultati

In questa sezione verranno riportati i risultati ottenuti dai test statistici *ANOVA 1-way*, *Tukey HSD pairwise* e *Tukey HSD multiple comparison* che sono stati effettuati grazie all'utilizzo dello script sviluppato.

Dal test dell'ANOVA 1-way sono stati ottenuti i risultati riportati in *Tabella 2*. Tali valori permettono di affermare che i 4 sistemi presi in esame assumono la stessa media, in quanto falliamo nel rifiutare la *Null Hypothesis* (valore di soglia  $\alpha$  posto a 0.05).

	Average Precision	P(10)	Rprec
F value	0.26982242831143294	0.35778395335621926	0.3508494180537509
P value	0.8471081879709074	0.7835600894153874	0.7885747493027742

Tabella 2. ANOVA 1-way

Il *Tukey HSD test* considerato in modalità *pairwise* permette di avere un confronto tra ogni coppia di sistemi. Da questo test si sono ottenuti i risultati riportati in *Tabella 3* per quanto riguarda l'analisi effettuata con la misura *Average precision*; si osserva anche qui che i sistemi sono molto simili. Anche il test svolto con la *P(10)* e la *Rprec* ha fornito risultati analoghi (reperibili al link Github nei file *tukey\_HSD\_p10.txt* e *tukey\_HSD\_rprec.txt*).

Multiple Comparison of Mean – Tukey HSD, FWER=0.05					
group 1	group 2	meandiff	lower	upper	reject
BM25	BM25_porter	-0.0018	-0.0877	0.0842	False
BM25	TF_IDF	-0.0005	-0.0865	0.0855	False
BM25	TF_IDF_not	-0.0251	-0.1111	0.0609	False
BM25_porter	TF_IDF	0.0012	-0.0848	0.0872	False
BM25_porter	TF_IDF_not	-0.0233	-0.1093	0.0626	False
TF_IDF	TF_IDF_not	-0.0246	-0.1106	0.0614	False

Tabella 3. Tukey HSD pairwise test con Average Precision

Per quanto riguarda il *Tukey HSD test*, considerato in modalità *multiple comparison*, è stato constatato nuovamente che i quattro sistemi risultano molto simili tra loro.

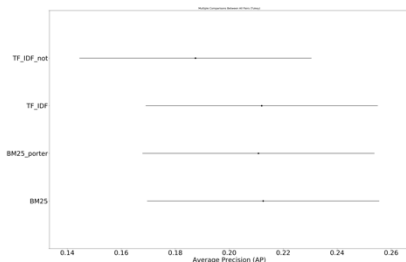


Figura 1. Tukey HSD test Average precision

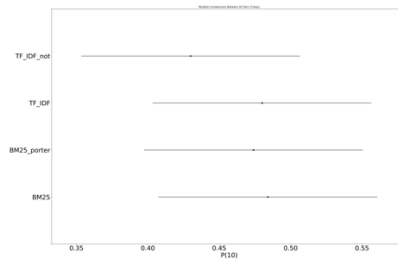


Figura 2. Tukey HSD test P(10)

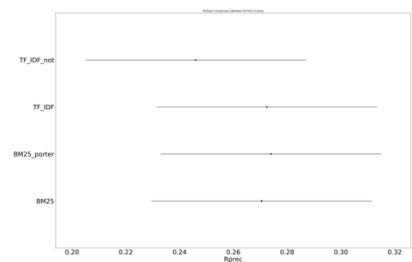


Figura 3. Tukey HSD test Rprec

Dalle figure riportate (Figure 1-3) si può notare che esiste sempre un'intersezione consistente tra gli intervalli di confidenza dei vari sistemi. È possibile osservare inoltre come il *TF\_IDF\_not* assuma valori leggermente più bassi rispetto agli altri sistemi in ognuna delle misure considerate e che, nell'analisi effettuata con la *P(10)*, anche il *BM25\_stem* tende a distaccarsi leggermente dagli altri. Inizialmente il test era stato svolto solo sulla misura *Average precision*, ma, notando la notevole somiglianza tra i sistemi, si è scelto di effettuarlo anche con la *P(10)* e la *Rprec* per individuare eventuali differenze. Osservando i risultati si può concludere che in tutti i test svolti si fallisce nel rifiutare la *Null Hypothesis*, quindi non sono state rilevate differenze sostanziali tra i sistemi analizzati.

In Figura 4 infine viene riportato il plot relativo al valore della *MAP* per ogni sistema. Si può notare che i valori del *TF\_IDF\_not* sono inferiori rispetto agli altri. Sono state inoltre rappresentate, in un istogramma per ogni sistema, le misure *P(10)* e *Rprec* per ogni topic (plot reperibili al link *Github* nei file "*nome\_misura\_nome\_sistema.png*"). Da questi si può notare che la *P(10)* assume valori più elevati rispetto all'*Average Precision* ed alla *Rprec*, diversi topic infatti assumono valori di precisione 1 in confronto allo 0.7 massimo ottenuto con le altre due misure. Possiamo quindi dedurre che i sistemi hanno maggiore precisione nel caso in cui si considerino i dieci documenti più rilevanti.

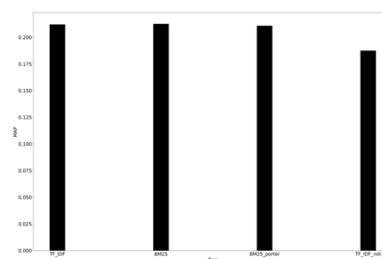


Figura 4. MAP

### 3. Conclusioni

Dalle analisi effettuate e dai risultati ottenuti si deduce che i sistemi che non fanno uso della *stop list* e del *Porter stemmer* forniscono risultati peggiori rispetto a quelli in cui vengono usate. Confrontando il *TF\_IDF\_not* e il *BM25\_porter* si evidenzia che il non utilizzo della *stop list*, accompagnato dall'applicazione del *Porter stemmer*, porta ad un leggero peggioramento, ma non così sostanziale come nel caso in cui non venga fatto utilizzo della rimozione delle *stopword*.

Queste analisi sono riassunte nella Tabella 4 dove vengono riportati i valori riassuntivi delle misure analizzate.

Nome sistema	MAP	P(10)	Rprec
BM25	0.2126	0.4840	0.2705
TF_IDF	0.2120	0.4800	0.2725
BM25_porter	0.2108	0.4740	0.2740
TF_IDF_not	0.1875	0.4300	0.2460

Tabella 4. Misure riassuntive

Si noti come i valori di tutte le misure siano nettamente più bassi nel sistema *TF\_IDF\_not*, mentre il *BM25\_stem* risulta molto vicino agli altri due sistemi. Dai risultati ottenuti si può concludere che il *TF\_IDF* e il *BM25* sono i sistemi con maggior precisione e affidabilità. Per un'analisi più accurata sarebbe necessario effettuare gli stessi test basandosi su statistiche ulteriori fornite da altri sistemi che utilizzino alternativamente il *Porter stemmer* o la *stop list* indipendentemente dal modello utilizzato; così facendo si potrebbe capire se effettivamente le prestazioni negative siano dovute all'assenza della *stop list* o del *Porter stemmer*.

Dal punto di vista dell'efficienza si può invece concludere che il non utilizzo della rimozione delle *stopword* porta a tempi di esecuzione maggiori e quindi il *TF\_IDF\_not* risulta essere il peggior sistema sia dal punto di vista dell'efficacia, che dell'efficienza.

### 4. Riferimenti

- [1] <http://terrier.org/>
- [2] [https://github.com/usnistgov/trec\\_eval](https://github.com/usnistgov/trec_eval)
- [3] <https://www.jetbrains.com/pycharm/>
- [4] <https://docs.python.org/3/library/os.html?highlight=os#module-os>
- [5] <https://matplotlib.org/>
- [6] <https://www.statsmodels.org/stable/index.html>
- [7] <https://docs.scipy.org/doc/>