Moratello Matteo 1205720
Corso di Laurea Magistrale in Ingegneria Informatica
A.A 2018/19 Corso di Reperimento dell'Informazione

Relazione Homework 1

Link al repository

https://github.com/Morat96/HW1_InformationRetrival

Specifiche

Per lo svolgimento dell'homework si è fatto uso delle seguenti specifiche:

- Terrier IR Platform v4.4 per *indicizzazione* ed esecuzione delle run sulla collezione TIPSTER;
- Terrier jtreceval (trec_eval) per la *valutazione* delle run;

Python 3 con le seguenti librerie:

- matplotlib per la realizzazione dei grafici e pandas per la realizzazione delle tabelle;
- numpy, scipy e statsmodels per il calcolo dei coefficienti del *test statistico* ANOVA 1-way e Tukey's HSD test;
- Jupyter notebook.

Procedura per indicizzazione, reperimento e valutazione

La prima fase del lavoro consiste nel settare terrier con i parametri specifici per il problema considerato. Per prima cosa si crea il file *collection.spec*, che raccoglie i percorsi di tutti i file della collezione, tramite l'istruzione *bin/trec_setup.sh/path/*.
Quindi si sono modificate le specifiche del file terrier.properties secondo i modelli specificati dalla consegna, le principali sono le seguenti:
termpipelines = Stopwords, PorterStemmer (per impostare le fasi di indicizzazione)
Impostazioni specifiche per le query:

TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=NUM
TrecQueryTags.process=TITLE,DESC
TrecQueryTags.skip=NARR

Inoltre, sono stati specificati i file contenenti i topic e qrels ed il modello considerato (una copia del file properties per ogni run è presente nella repository).
Tramite il comando *bin/trec_terrier.sh -i* si crea l'indice. Quindi si valuta la run eseguendo i comandi *bin/trec_terrier.sh -r* e *bin/trec_terrier.sh -e -p*.
Si ottengono quindi i file della run ed il file *.eval* , nel quale si trovano le misure di valutazione (in particolare MAP,Rprec e Precision at 10) calcolate sia nella media delle query che per le singole query.

Risultati sperimentali

Confronto valori di MAP, Rprec e Precision at 10

	MAP	Rprec	Precision at 10
BM25 Stoplist,PorterStemmer	0.2125	0.2705	0.482
TF*IDF Stoplist,PorterStemmer	0.2123	0.2725	0.478
BM25 No Stoplist,PorterStemmer	0.1245	0.1701	0.302
TF*IDF No Stoplist,No PorterStemmer	0.1876	0.2485	0.426

Figura 1 Risultati di MAP, Rprec e Precision at 10 delle diverse run

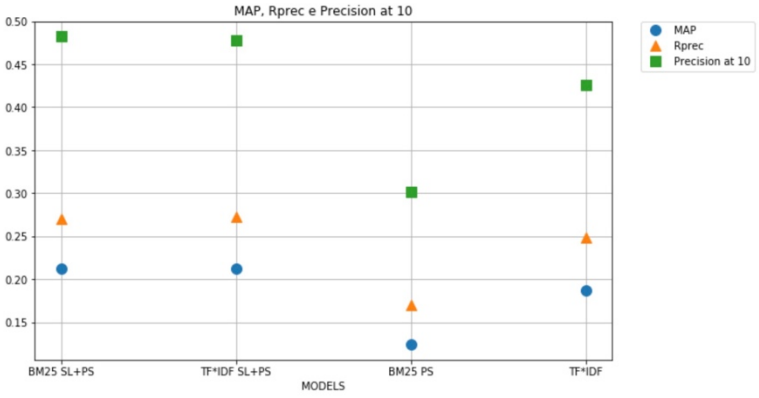
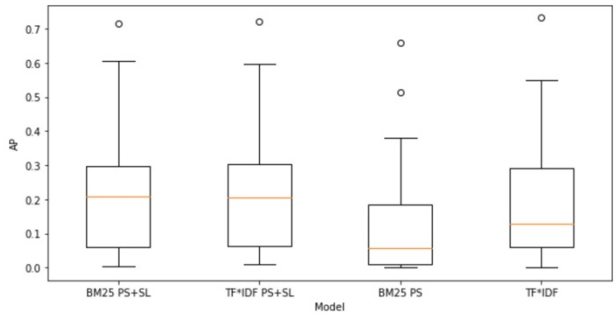
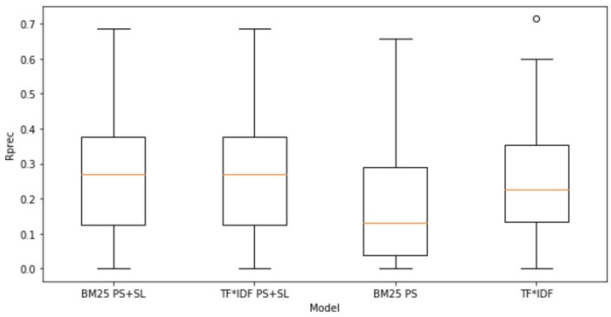


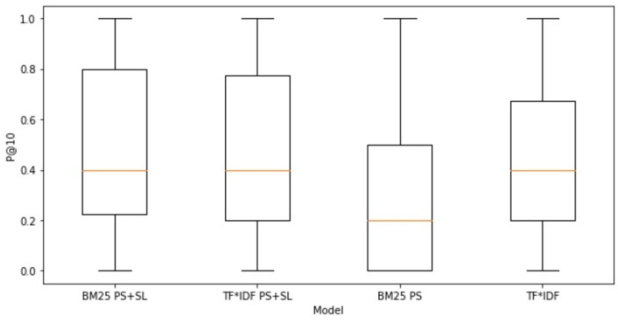
Figura 2 Grafico con risultati di MAP, Rprec e Precision at 10

Come si può vedere dalla tabella (Figura 1) e dal grafico (Figura 2), le run con i risultati migliori sono le prime due, cioè quelle eseguite con i sistemi BM25 e TF*IDF con Porter Stemmer e Stoplist.
Nella collezione considerata, il processo di riduzione di una parola alla sua forma radice (stemming) e l'esclusione dall'indice di parole molto frequenti, quindi poco significative nella fase di reperimento (stoplist), risulta molto efficace indipendentemente dal modello considerato (BM25 e TF*IDF), essenzialmente i valori trovati sono equiparabili. In questo caso attuare le due fasi dell'indicizzazione comporta un effettivo miglioramento a discapito di un costo computazionale più elevato. I risultati peggiori sono invece ottenuti dalla run 3 eseguita con modello BM25 con PorterStemmer.

BoxPlot



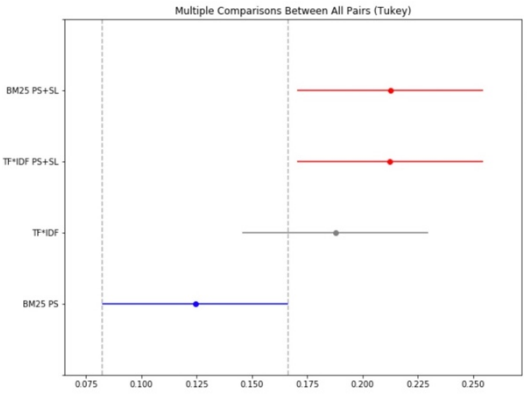


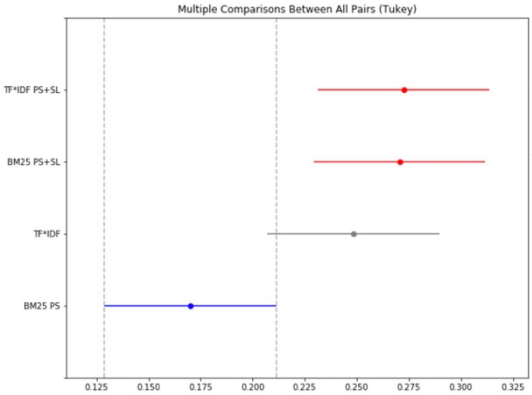


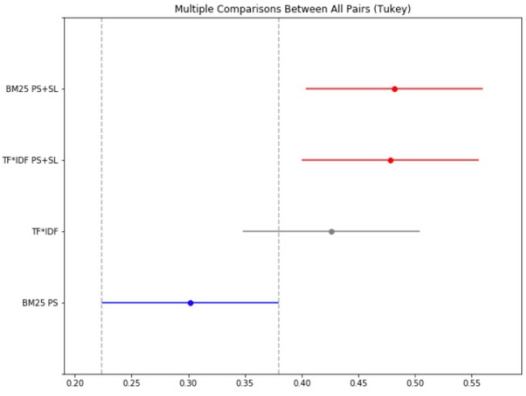
I Boxplot mostrano le distribuzioni delle misure (AP, Rprec e Precision at 10) per ciascun sistema considerato.
Dai plot si ha una conferma della migliore prestazione avuta dai modelli BM25 e TF*IDF con Stemmer e Stoplist. Il peggiore risulta invece il modello BM 25 solo Stemmer, la sua distribuzione per ogni misura è concentrata vicino allo zero, molti valori sono infatti nulli (il sistema non ha reperito alcun documento).
Per evidenziare una differenza statistica tra le run considerate si ricorre al test ANOVA 1-way.

ANOVA 1-way e Tukey's HSD Test

Il procedimento completo del calcolo dei test è riportato nel notebook.







I plot mostrano il risultato del Tukey's Test rispettivamente per le misure AP, Rprec e Precision at 10.

Il test ANOVA permette di verificare se due o più sistemi sono statisticamente uguali, partendo dall'ipotesi nulla:
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
Il test evidenzia che la null hypothesis viene rifiutata in quanto PR(>F) < α per ciascuna delle misure di valutazione, si può affermare quindi che almeno un modello ha una distribuzione con una differenza statisticamente significativa rispetto ad un altro modello. Attraverso il test di Tukey è stato possibile identificare i modelli appartenenti al *top group* ed ha evidenziato che il sistema **BM25 PorterStemmer** presenta differenze significative con gli altri tre modelli. Possiamo concludere che i modelli appartenenti al *top group* sono quelli delle prime due run, cioè **BM 25 PorterStemmer e Stoplist** e **TF*IDF PorterStemmer e Stoplist.**