Christian Obetti 1205719
Corso di Laurea Magistrale in Ingegneria Informatica
A.A. 2018/19 Corso di Reperimento dell'informazione

Relazione Homework 1

Specifiche

Per lo svolgimento delle attività previste dall'homework si è fatto uso di:

- Terrier IR Platform v4.4 per indicizzazione ed esecuzione delle diverse run sulla collezione TIPSTER;
- Python3;
- Librerie numpy, scipy e statsmodels per il calcolo dei coefficienti dei test ANOVA one-way e Tukey HSD;
- Librerie matplotlib per la realizzazione dei grafici e pandas per la realizzazione delle tabelle;
- Jupyter Notebook

Procedimento

Nella prima fase si è fatto uso di terrier per creare il file *collection.spec* che raccoglie i percorsi di tutti i file della collezione, tramite il comando da terminale *bin/trec_setup.sh /path/*. Quindi si sono modificate le specifiche del file *terrier.properties* seguendo le istruzioni della consegna (una copia delle properties per ogni run è presente nella repository). Tramite il comando *bin/trec_terrier.sh -i* si è dunque creato l'indice. Successivamente, tramite il comando *bin/trec_terrier.sh -r*, si è eseguita la run. I file contenenti i topics e le query rilevanti sono state specificate nel file *.properties*, così come il modello usato. Infine, si sono calcolati i coefficienti necessari per la valutazione (in particolare MAP, RPrecision e Precision at 10) tramite l'istruzione *bin/trec_terrier.sh -e -p*. I coefficienti sono stati scritto da terrier in un file *.eval*.

Nella seconda fase si sono estratti i coefficienti di interesse e visualizzati in modo più ordinato in un Jupyter Notebook. Si è quindi condotto il test ANOVA one-way e il test Tukey HSD per una migliore comprensione dei risultati.

Risultati sperimentali

Confronto fra le precision

Run	MAP	RPrecision	Precision at 10
1.0	0.2125	0.2705	0.482
2.0	0.2123	0.2725	0.478
3.0	0.1245	0.1701	0.302
4.0	0.1876	0.2485	0.426

Figura 2 Tabella riassuntiva con tutti i risultati delle diverse run

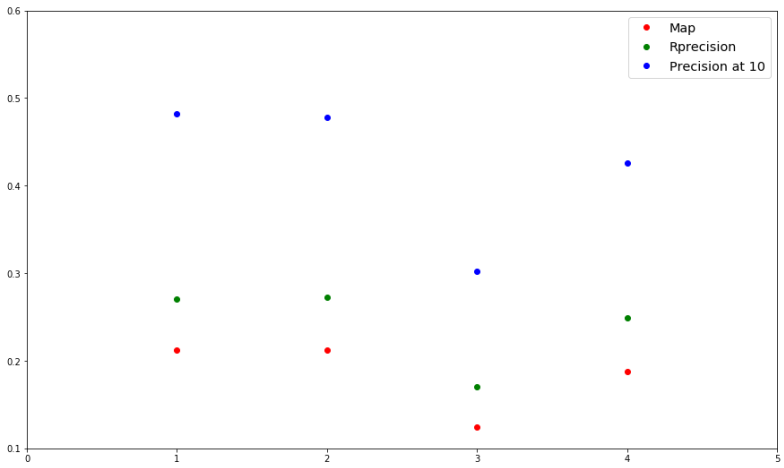

Figura 1 Plot che individua graficamente le differenze nei risultati

Come si può osservare, le run che hanno ottenuto i migliori risultati sono state le prime due, che facevano entrambe uso di una stoplist e di un porter stemmer. I risultati peggiori invece si sono ottenuti nella terza run, usando il modello BM25 senza l'ausilio di una stoplist, e confrontandoli con la quarta run, in cui non si

è fatto uso né della stoplist, né di un porter stemmer, si evidenzia come il modello IF*IDF, che ordina i termini secondo la loro frequenza nei documenti, è risultato migliore.

Test statistico

Si consideri il boxplot delle diverse distribuzioni basate sui valori di *Mean Average Precision*:

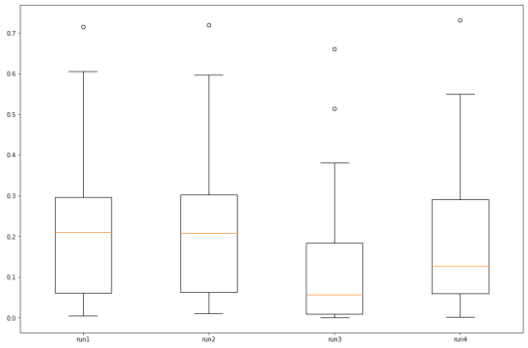

Figura 3 Boxplot delle distribuzioni nelle run

Come si può evincere dal grafico, le distribuzioni non sono molto diverse fra loro nelle dimensioni, tranne che per la run 3.

Gli elementi mediani rappresentano i valori di MAP di ciascuna delle run e hanno lo stesso significato descritto in precedenza.

Si intuisce che le run 1 e 2 siano diverse dalla run 3, mentre per la run 4 si notano risultati intermedi, con una distribuzione molto simile alle prime due. Per verificare l'effettiva differenza fra le distribuzioni occorre svolgere il test ANOVA one-way.

ANOVA one-way e Tukey HSD

Nel notebook si è descritto il procedimento attuato. Qui si evidenziano una volta in più le stesse osservazioni fatte nel notebook.

Il test ANOVA permette di verificare se l'ipotesi nulla, ovvero che le quattro distribuzioni siano congruenti, si verifichi. Il test evidenzia un valore di P di all'incirca 0.02, inferiore alla soglia di confidenza α=0.05. Quindi l'ipotesi nulla è violata e almeno una delle distribuzioni è diversa da un'altra. Per determinare quali, occorre eseguire il test Tukey HSD, il quale effettua un confronto a coppie e stabilisce quali distribuzioni hanno violato l'ipotesi nulla. Di seguito è mostrato il plot.

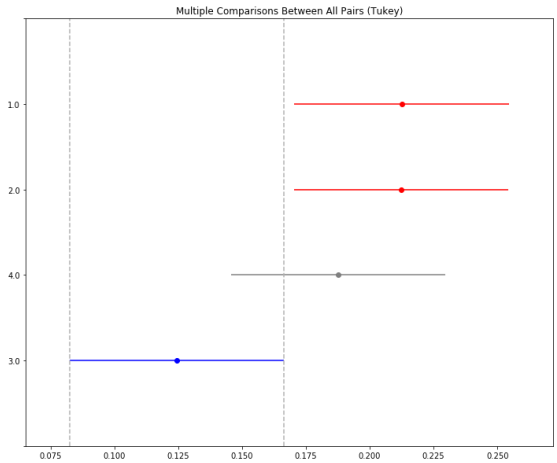

Figura 4 Plot dei risultati del test Tukey HSD

Il test indica che l'ipotesi nulla è violata nel confronto fra distribuzioni di run 1 e 3 e run 2 e 3, il che ci suggerisce con esattezza che la run 3 non appartiene al "top group". Per quanto riguarda la run 4, le differenze non sono così grandi né da sottolineare una differenza con la run 3 o con le run 1 e 2. Tuttavia, come affermato in precedenza, la precisione media è inferiore e quindi le prestazioni effettivamente peggiori di quelle delle run 1 e 2.

Il test statistico è stato condotto anche per i valori di R-Precision e Precision at 10. Tuttavia, non sono emerse particolari differenze rispetto a quello basato su MAP. Si veda il commento sul notebook per approfondimento.

Link repository

https://github.com/chrisob194/informationretrievalhw