Information retrieval - Homework 1

Paolo Bonato mat:1156640

January 3, 2019

1 Introduzione

Il compito consiste nell'esecuzione di diverse run di un sistema di reperimento e la valutazione dei risultati ottenuti utilizzando la collezione sperimentale TREC7, 50 topic e un pool con 2 gradi di rilevanza: R, NR. Le specifiche del materiale utilizzato:

- Sistema di reperimento: Terrier versione 4.4
- Linguaggio di programmazione e compilatore: Python versione 2.7.12
- Libreria per la produzione di grafici: matplotlib versione 2.2.3
- Librerie per le analisi statistiche: SciPy versione 1.1.0, statsmodels versione 0.9.0

Link alla repository GitHub destinata alla consegna:
https://github.com/pbonato3/IR_HW1_18-19

2 Svolgimento

In seguito ci si riferir alle run come: **Run 0:** Stoplist, PortStemmer, modello BM25, **Run 1:** Stoplist, PortStemmer, modello TF*IDF, **Run 2:** No stoplist, PortStemmer, modello BM25, **Run 3:** No stoplist, No Stemmer, modello TF*IDF.
Per ogni run stata eseguita una diversa indicizzazione e due test di reperimento, ovvero con e senza il meccanismo di query expansion. I tag considerati per il retrieval sono sia "TITLE" che "DESC", ignorando invece il tag "NARR". Infine sui risultati stato lanciato il comando di valutazione *trec_eval* per ottenere le metriche rispetto a tutte le query (riportate nella repository)

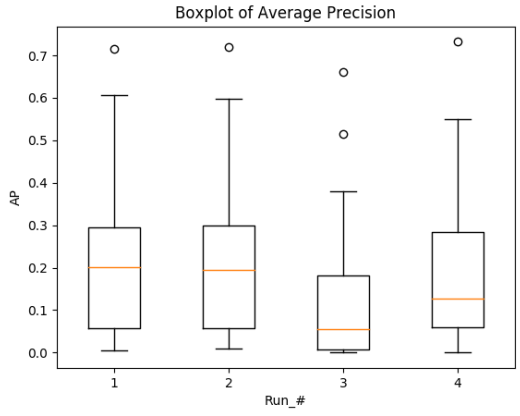
Le metriche che andremo a confrontare sono l'Average Precision (**AP**), la precision con cut-off ai primi 10 documenti reperiti (**P@10**) e la precision con cut-off al numero di documenti rilevanti (**RPrec**). Le medie delle metriche ottenute sono:

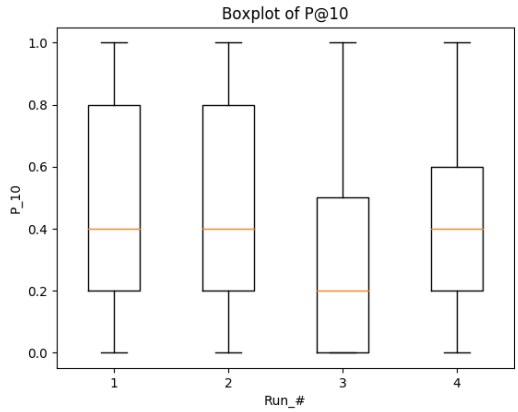
	Run_0	Run_1	Run_2	Run_3
MAP:	0.2125	0.2123	0.1245	0.1876
P@10:	0.482	0.478	0.302	0.426
RPrec:	0.2705	0.2725	0.1701	0.2485

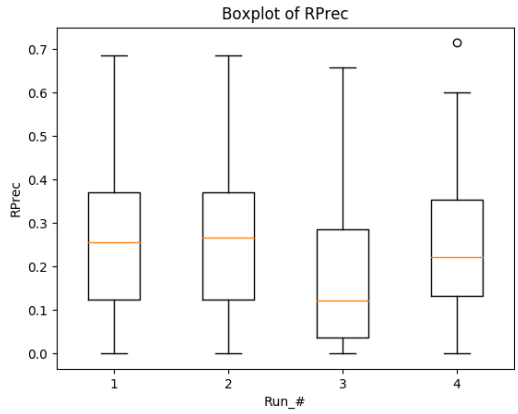
Utilizzando il meccanismo di Query Expansion i risultati sono i seguenti:

	Run_0	Run_1	Run_2	Run_3
MAP:	0.2538	0.2512	0.1697	0.2276
P@10:	0.508	0.514	0.362	0.428
RPrec:	0.2936	0.2934	0.2033	0.2769

Come ci si poteva aspettare i risultati delle run che utilizzano Stoplist e PortStemmer sono migliori delle altre. Questo si puó osservare anche dai boxplot delle metriche per ogni query:







In particolare risulta evidente che la run numero 2, con sistema BM25, PortStemmer ma senza Stoplist, ha una media inferiore alle altre run. Per capire per se sia una differenza statisticamente significativa occorre condurre un test ANOVA 1-way sulle run. Viene qui riportato il test riguardante le tre metriche principali che é stato condotto sia sulle quattro run assieme che sulle coppie che utilizzano lo stesso modello:

```
ANOVA 1-way, AP, all runs:
F_value:  3.3049517787341163  P_value:  0.02137066541175464
ANOVA 1-way, AP, BM25 runs:
F_value:  7.824921045638202  P_value:  0.006226581737823212
ANOVA 1-way, AP, TF_IDF runs:
F_value:  0.5364900435963  P_value:  0.46567557019319616


ANOVA 1-way, P@10, all runs:
F_value:  3.9664133596022144  P_value:  0.008990650615442174
ANOVA 1-way, P@10, BM25 runs:
F_value:  8.803151745686728  P_value:  0.003795414427067481
ANOVA 1-way, P@10, TF_IDF runs:
F_value:  0.8512625894030074  P_value:  0.3585084710269246


ANOVA 1-way, RPrec, all runs:
F_value:  4.644451217758076  P_value:  0.003695696907257868
ANOVA 1-way, RPrec, BM25 runs:
F_value:  10.114080275211915  P_value:  0.0019816508846238735
ANOVA 1-way, RPrec, TF_IDF runs:
F_value:  0.5622715036511406  P_value:  0.4551801872976361
```

Volendo confrontare tutte le possibili coppie di run é stato eseguito un test di Tukey di cui viene ripostato il risultato per le medie delle Average Precision:

```
 Multicomare of AP metrics
Multiple Comparison of Means - Tukey HSD,FWER=0.05
=============================================
group1 group2 meandiff  lower   upper  reject
---------------------------------------------
Run_0  Run_1  -0.0003  -0.0845  0.084  False
Run_0  Run_2  -0.0886  -0.1729 -0.0044  True
Run_0  Run_3  -0.0247  -0.1089  0.0596 False
Run_1  Run_2  -0.0884  -0.1727 -0.0041  True
Run_1  Run_3  -0.0244  -0.1087  0.0599 False
Run_2  Run_3   0.064   -0.0203  0.1483 False
---------------------------------------------
```

Si é infine scelto di escludere la run 2 e ripetere il test Anova:

```
ANOVA 1-way, AP, without Run_2:
F_value:  0.3584143491005936  P_value:  0.6994050671254


ANOVA 1-way, P@10, without Run_2:
F_value:  0.612133782732694  P_value:  0.5435974429063013


ANOVA 1-way, RPrec, without Run_2:
F_value:  0.34858024177755353  P_value:  0.7062830871525674
```

3 Conclusioni

Dai risultati ottenuti si puó concludere che le diverse run hanno delle variazioni nelle metriche di valutazione che sono statisticamente significative e in particolare si puó rifiutare la null hypotesis con $\alpha = 0.5$.
La run che si distingue é la numero 2 come dimostrato dai test Anova e di Tukey e una volta esclusa non piú possibile evidenziare differenze statisticamente significative. Da questo si evince che il modello BM25 in assenza di Stoplist ha una performace peggiore del modello TF_IDF, anche se quest'ultimo opera in assenza di stemmer.
Dall'esperimento condotto non é invece possibile dimostrare un'evidete differenza tra i due modelli quando essi utilizzano sia Stoplist che PortStemmer.
Utilizzando il meccanismo di QueryExpansion si ottiene un sensibile miglioramento dei risultati che rende meno evidente il distacco tra le medie, ma nonostante questo si giunge alle medesime conclusioni. I dati a riguardo sono disponibili nella repository.