RELAZIONE Homework 1

REPERIMENTO DELL'INFORMAZIONE

Barisan Anna
Matricola: **1206600**
Corso di Laurea Magistrale in Ingegneria Informatica

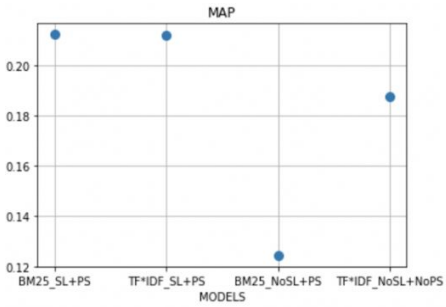
Sistema di IR utilizzato: Terrier IR Platform v4.4;
Libreria di valutazione: implementazione in Java di Terrier di trec_eval;
Specifiche utilizzate per la valutazione: Python 3, Jupyter Notebook, librerie *numpy*, *scipy* e *statsmodels* (calcolo dei coefficienti dei test ANOVA one-way e Tukey HSD), librerie *pandas* e *matplotlib* per la realizzazione delle tabelle e dei grafici;

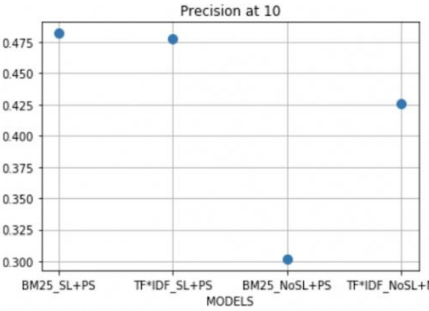
Per la creazione dell'indice si è proceduto in questo modo: per la creazione della collezione relativa al file "*collection.spec"* è stato utilizzato il comando *sh bin/trec_setup.sh /path/* (utilizzando il percorso relativo alla cartella contenente i documenti della collezione TIPSTER); dopo aver modificato il file "*terrier.properties*" (una copia per ogni sistema è presente nella repository) per impostare le diverse fasi dell'indicizzazione richieste dalla consegna (rimozione delle stopword e stemming) specificate nella riga *termpipelines=Stopwords,PorterStemmer,* e impostato le specifiche per le query (*TrecQueryTags),* si è proceduto alla effettiva creazione dell'indice attraverso il comando *sh bin/trec_terrier.sh -i .*
Infine i due comandi successivi specificano le interrogazioni e i rispettivi giudizi di rilevanza utilizzati per la valutazione: *bin/trec_terrier.sh -r -Dtrec.model=BM25 (oppure TF_IDF) -Dtrec.topics=/path_to_topics bin/trec_terrier.sh -e -p -Dtrec.qrels=/path_to_qrels.trec7.txt.* Questi comandi sono stati ripetuti per i 4 sistemi. A questo punto otteniamo i file delle run e il file .eval nei quali si trovano i valori delle misure calcolate per ogni singola query e per tutte le query insieme. I valori ottenuti sono:

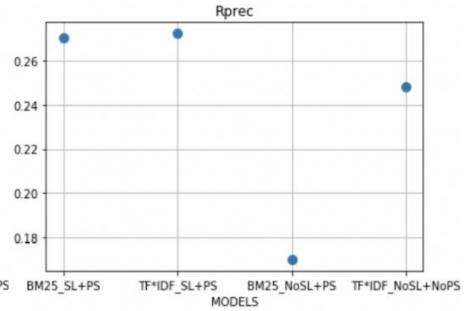
	MAP	Rprec	Precision@10	
BM25_Stoplist_PorterStemmer	0.2125	0.2705	0.482	BM25_SL+PS
TF*IDF_Stoplist_PorterStemmer	0.2123	0.2725	0.478	TF*IDF_SL+PS
BM25_NOStoplist_PorterStemmer	0.1245	0.1701	0.302	BM25_NoSL+PS
TF*IDF_NOStoplist_NOPorterStemmer	0.1876	0.2485	0.426	TF*IDF_NoSL+NoPS

Dai valori ricavati si può dedurre a prima vista che i sistemi più performanti sono i primi due: BM25 e TF*IDF che utilizzano entrambi sia la rimozione delle stopword sia lo stemming (i valori si discostano di poco). Questo a significare che su questa collezione l'utilizzo di tecniche che riducono la dimensione dell'indice, eliminando parole non portatrici di significato, e che raggruppano parole semanticamente uguali, migliora comprensibilmente le performance e la precisione. In questo caso spendere risorse, attività e calcoli per attuare queste fasi dell'indicizzazione è vantaggioso e comporta un effettivo miglioramento. Tra i quattro, il sistema che risulta peggiore è BM25 con l'utilizzo dello stemmer ma senza la rimozione delle stopword.
Di seguito vengono riportati i plot di MAP, Rprec e Precision at 10 relativi ai quattro sistemi.

L'analisi statistica è stata eseguita in uno jupiter notebook: è stato eseguito un test ANOVA 1-way per ogni distribuzione MAP, Rprec, Precision at 10. Esse hanno mostrato risultati praticamente identici, confermando le ipotesi fatte inizialmente. Viene di seguito analizzata più dettagliatamente l'analisi sui valori MAP, esemplificativa rispetto alle altre due, che sono state meglio discusse all'interno del notebook.

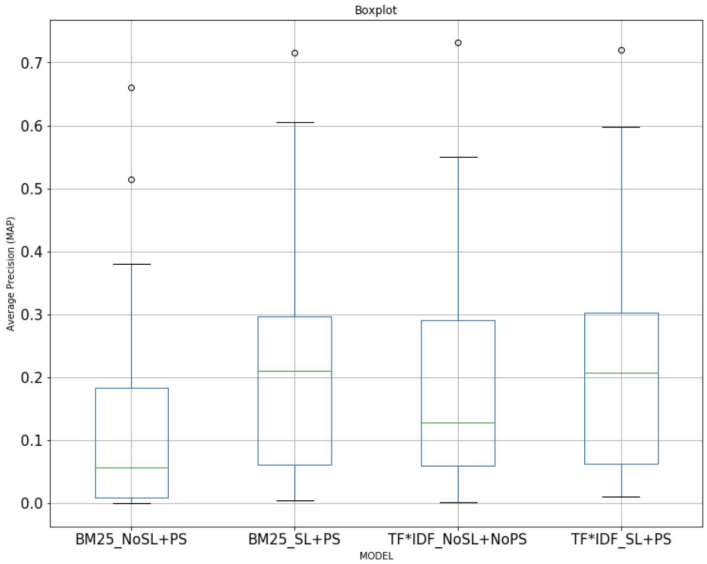




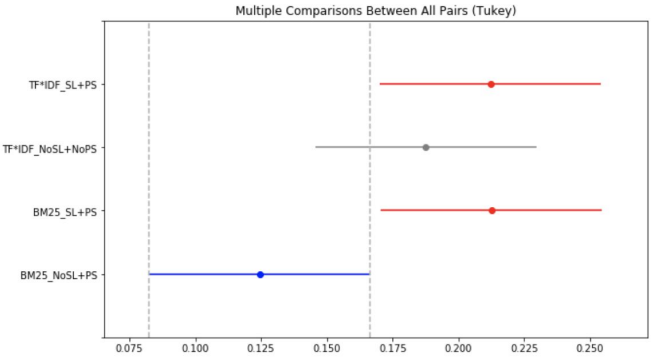


Il grafico dei boxplot relativi alle MAP mostra che le distribuzioni non sono molto diverse fra loro nelle dimensioni, ad eccezione di "BM25_NoSL+PS" che risulta la peggiore in termini di risultati; gli elementi mediani rappresentano i valori di MAP di ciascuna run e anche qui si intuiscono le diversità sopra citate.
Il test ANOVA mostra la stessa differenza tra i sistemi. E' risultato possibile rifiutare la null hypothesis (ovvero che le quattro distribuzioni sono congruenti) poichè il valore trovato pvalue ≤ α (0.02214 ≤ 0.05): si può affermare che almeno un modello ha una media con differenza statisticamente significativa rispetto ad un altro ma non possiamo ancora dire quale, è necessario confrontare ogni coppia di modelli per determinare quali sono quelli significativamente differenti. E' stato quindi eseguito il test di Turkey utilizzando la multiple comparison che ha indicato le coppie di modelli diversi:
"BM25_NoSL+PS" - "BM25_SL+PS" e "BM25_NoSL+PS" - "TF*IDF_SL+PS" (come si era previsto osservando solo le misure calcolate inizialmente). I plot seguenti esemplificano ciò.

Possiamo concludere quindi, anche basandoci sugli altri due test, che i sistemi appartenenti al top group sono i primi due, "BM25_SL+PS" e "TF*IDF_SL+PS"; essi sono significativamente diversi dal terzo modello "BM25_NoSL+PS" che non appartiene a questa categoria avendo risultati peggiori. Il quarto modello "TF*IDF_NoSL+NoPS" si colloca nel mezzo: non è diverso dai primi due, però si intuisce una differenza, tenendo presente che esso non risulta significativamente diverso dal terzo sistema e che la sua precisione media è inferiore e le prestazioni peggiori.



Test di Turkey sui valori MAP



Link al repository:
`https://github.com/annabarisan/Homework1_IR_BarisanAnna.git`
Sono presenti i file delle run e i file di valutazione dei sistemi (divisi in 4 cartelle separate, una per modello) e il file jupyter notebook ("Evaluation_ HW1_IR.ipynb") utilizzato per la valutazione: esso contiene ulteriori commenti più approfonditi, i plot e il test statistico applicato alle tre misure.