Andrea Grendene Matricola: 1184985

Reperimento dell'informazione - Homework 1

Strumenti utilizzati e link alla repository

- Terrier versione 4.4 per l'indicizzazione e il reperimento
- Trec_eval versione 9.0.4 per il calcolo delle unità di misura per la valutazione
- Matlab versione R2018b (9.5) per l'esecuzione del test statistico ANOVA e il test di Tukey HSD
- Repository: https://github.com/Cenze94/IR_HW1_Andrea_Grendene

Scelte implementative

Oltre alle 4 run richieste per lo svolgimento dell'homework è stato deciso di aggiungerne altre 3, in cui al posto dello stemmer Porter è stato usato il Weak Porter, mentre le altre impostazioni sono state lasciate invariate. Le proprietà del file 'terrier.properties' sono state lasciate inalterate per ogni run, ad eccezione ovviamente di 'trec.model' per la scelta del modello, in cui sono stati usati TF*IDF e BM25, di 'termpipelines' per la scelta dell'utilizzo o meno di stop list e stemmer e per indicare quale stemmer usare, e di 'terrier.index.path' per indicare dove salvare i file di output dell'indicizzazione. Per quanto riguarda i parametri di BM25 sono stati usati i valori di default di Terrier, ossia k1=1.2 e b=0.75. Invece le impostazioni invariate più importanti sono la stop list usata, ossia quella standard fornita da Terrier, le sezioni dei topic usate per effettuare il reperimento, che in questo caso viene usato solo il 'title', mentre 'description' e 'narrative' vengono ignorati, 'ignore.low.idf.terms', ovvero se considerare o meno i termini con IDF basso, che è stato mantenuto a 'false' e quindi nessun termine viene ignorato, 'tokeniser' e 'trec.encoding', ossia le caratteristiche relative all'analisi lessicale, impostati rispettivamente alla tokenizzazione per la lingua inglese e alla codifica in UTF-8, dato che i documenti sono tutti in inglese e scritti in tale codifica.
Per semplicità le run sono state identificate con un numero, basato sull'ordine delle run richieste nella consegna dell'homework; perciò 'Run 1' corrisponde a quella che utilizza la stop list, lo stemmer Porter e BM25, 'Run 2' a quella che usa la stop list, lo stemmer Porter e TF*IDF, 'Run 3' a quella che utilizza lo stemmer Porter e BM25 e 'Run 4' a quella che usa soltanto TF*IDF. Per identificare le run che usano lo stemmer Weak Porter è stata aggiunta la sigla 'WPS' al nome.

Struttura della repository

All'interno delle cartelle 'Run1_Stopword+Stemmer+BM25', 'Run2_Stopword+Stemmer+TF_IDF', 'Run3_Stemmer+BM25' e 'Run4_TF_IDF' sono contenuti i file relativi ai risultati ottenuti con Terrier. Invece all'interno delle cartelle 'map_charts', 'P@10_charts' e 'Rprec_charts' sono contenuti i grafici e il codice Matlab per l'esecuzione del test statistico ANOVA. Sono presenti poi due file di script in Python, 'Extract_values.py' e 'Extract_values_weak_porter_stemmer.py': entrambi servono per estrarre e suddividere i valori di MAP, P@10 e Rprec di tutte le run, salvandoli nei file in formato '.txt' contenuti nella stessa directory. Successivamente questi file sono stati caricati in Matlab tramite codice e salvati nei file in formato '.mat' sempre presenti nella stessa directory. Il primo file di script ottiene le misure per le 4 run richieste dall'homework, il secondo aggiunge le altre 3 run. All'interno delle cartelle delle Run sono presenti due sottocartelle, 'Standard' e 'Weak_Porter_Stemmer', dove sono contenuti i file relativi alla rispettiva run e usando il Weak Porter Stemmer o il Porter Stemmer; l'unica eccezione è 'Run4_TF_IDF', dove non è presente la cartella 'Weak_Porter_Stemmer'. All'interno di queste sottocartelle sono presenti due file, 'evaluation' e 'terrier. properties', e la cartella 'results': il primo contiene l'output di 'trec_eval', ossia tutte le misure ottenute per la rispettiva run, da cui verranno estratte quelle necessarie per il calcolo del test statistico ANOVA con gli script citati in precedenza; il secondo è il file delle impostazioni usate da Terrier per l'esecuzione della rispettiva run; la terza invece contiene i due file di output del reperimento dei documenti eseguito con Terrier. All'interno delle cartelle relative alle unità di misure sono contenute sempre le due sottocartelle 'Standard' e 'Weak_Porter_Stemmer', al loro interno sono presenti il codice Matlab per il calcolo del test statistico ANOVA e i file ottenuti come output, ossia due PDF contenenti i due grafici costruiti per rappresentare i risultati rispettivamente del test statistico e del test di Tukey HSD.

Risultati ottenuti

	Run 1	Run 2	Run 3	Run 4	Run 1 WPS	Run 2 WPS	Run 3 WPS
MAP	0.1828	0.1821	0.1854	0.1692	0.1776	0.1773	0.1815
P@10	0.4180	0.4200	0.4300	0.4060	0.4020	0.4020	0.4180
Rprec	0.2392	0.2391	0.2404	0.2288	0.2309	0.2313	0.2383

È stato deciso di non presentare in questa relazione la tabella e i grafici relativi soltanto alle prime 4 run, perché tali risultati sono già inclusi nella tabella e nei grafici riportati, mentre nella repository sono stati mantenuti i file di entrambe le versioni per facilitare il confronto con i risultati attesi dell'homework.
Analizzando i valori ottenuti si può notare subito che, ordinando le run in base ai risultati di una misura, si tende ad ottenere lo stesso ordine, a prescindere da quale misura è stata scelta. In altre parole dove ad esempio la MAP tende ad essere più bassa anche la Rprec e la P@10 tendono ad essere bassi e viceversa. Analizzando tale ordine si nota subito che la run peggiore è 'Run 4', ossia quella senza stemmer né stop list; tale risultato non è sorprendente, perché senza lo stemmer diventa impossibile identificare i casi in cui un certo termine viene usato in un'altra forma, come ad esempio il verbo 'to get' che al presente può diventare 'get' e 'gets' a seconda del soggetto. La run migliore invece è 'Run 3', ossia quella in cui non viene usata la stop list; anche tale risultato non dovrebbe sorprendere più di tanto, perché l'utilizzo della stop list è adottato per motivi di spazio occupato dall'indicizzazione e di velocità del reperimento, a scapito però della precisione di quest'ultimo. Tra 'Run 1' e 'Run 2', ossia tra quella che utilizza BM25 e quella che usa TF*IDF, si può notare che la migliore risulta essere la prima, anche se di poco, perciò in questo caso BM25 porta a risultati leggermente migliori di TF*IDF; l'unica eccezione è P@10, che risulta essere migliore in 'Run 2', anche se di molto poco, quindi in questo caso TF*IDF è leggermente più efficace nel reperire documenti rilevanti tra i primi 10 documenti reperiti. L'utilizzo dello stemmer Weak Porter porta a risultati peggiori rispetto alle run in cui viene usato Porter, quindi in questo caso l'utilizzo di uno stemmer meno aggressivo non è consigliato. Confrontando la run migliore che usa lo stemmer Weak Porter, ossia 'Run 3 WPS', con quella generalmente peggiore che usa lo stemmer Porter, ovvero 'Run 2', si può notare che la peggiore sia sempre la prima. Tale classifica è presente anche nei grafici ottenuti eseguendo il codice Matlab del test statistico ANOVA 1-way e del test di Tukey HSD, come si può notare nei seguenti esempi:

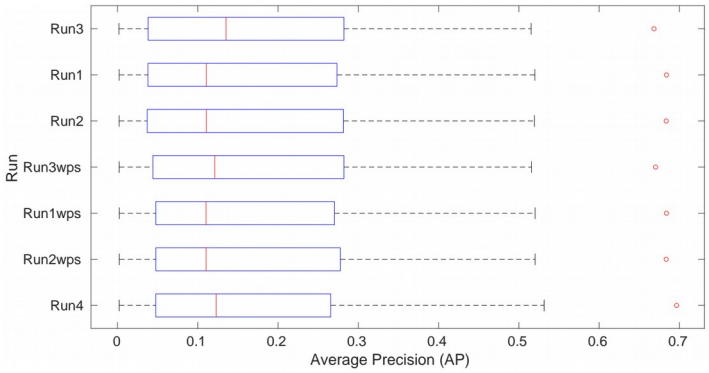

Figura 1

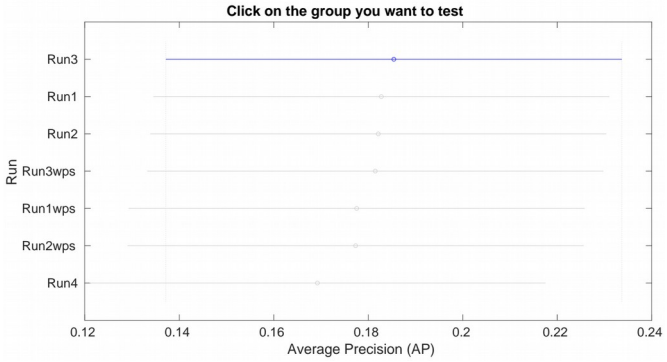

Figura 2

Sono stati riportati solamente i grafici relativi all'Average Precision di ogni singolo topic, dato che quelli relativi alle altre misure sono concordi e reperibili nella repository. Il test statistico ANOVA 1-way permette di verificare se l'ipotesi H0 di avere le medie di tutte le run uguali tra loro sia vera: in questo caso non c'è evidenza per rifiutare tale ipotesi, come si può notare nel grafico boxplot (Figura 1), in cui le mediane sono tutte vicine fra loro e quindi non c'è una significativa variazione fra i sistemi. Di conseguenza tutte le run appartengono al top-group, come risulta nel grafico del test di Tukey (Figura 2): lo scopo di tale test è l'individuazione dei gruppi con media significativamente diversa da quelle degli altri sistemi, dato che il test statistico ANOVA permette solo di capire se ce ne sono e non quali sono; dato che l'ipotesi H0 è valida allora non esistono tali gruppi.

Conclusioni

Con la collezione in analisi la configurazione migliore è risultata essere quella che utilizza lo stemmer Porter e non usa alcuna stop list. A seguire ci sono le configurazioni con la stop list, i sistemi che utilizzano lo stemmer Weak Porter e infine la configurazione senza stop list né stemmer. Tale risultato è conforme con le aspettative iniziali, dato che l'utilizzo dello stemmer ha portato a risultati migliori, l'assenza della stop list ha permesso di analizzare più termini, quindi di avere maggiore precisione e un risultato migliore, e l'utilizzo dello stemmer Weak Porter, meno aggressivo, al posto di Porter, più diffuso, ha portato ad un reperimento meno efficiente. Per quanto riguarda il confronto fra i modelli TF*IDF e BM25, il migliore è risultato essere il secondo, sebbene la differenza sia molto scarsa. Tra le varie run non ci sono grosse differenze per quanto riguarda il valore delle tre unità di misura prese in considerazione, e questo dato è stato confermato dagli esiti del test statistico ANOVA 1-way e del test di Tukey HSD, in cui non è emersa alcuna run la cui media di una qualunque unità di misura risultasse significativamente diversa rispetto alle altre.