Sistemi per il recupero delle informazioni
Laurea Magistrale in Editoria e Giornalismo
Prova scritta del 5 luglio 2012

Cognome e nome:_____ **Matricola:**_____

Domanda 1	Domanda 2	Domanda 3	Domanda 4	Domanda 5	Domanda 6	Totale

Istruzioni:

- È vietato portare all'esame libri, eserciziari, appunti e dispense. Chiunque venga trovato in possesso di documentazione relativa al corso, in formato analogico e/o digitale, – anche se non strettamente attinente alle domande proposte – vedrà annullata la propria prova.
- Scrivere solo sui fogli distribuiti, cancellando le parti di brutta con un tratto di penna. Non separare questi fogli. Non utilizzare la penna rossa. Scrivere nome e cognome su tutti i fogli utilizzati.
- Tempo a disposizione: 1 ora e 45 minuti.

1) Si rappresenti graficamente e si descriva l'architettura di un sistema per il recupero delle informazioni e in particolare quali sono gli scopi principali dell'indicizzazione.
2) Si descrivano cosa sono e come vengono utilizzati TEI e DocBook.
3) Si citino i diversi tipi di interrogazione messi a disposizione dai sistemi per il recupero delle informazioni. Si descrivano in particolare le interrogazioni fuzzy/approssimate, presentando brevemente le idee su cui si basa il recupero dei documenti per questo tipo di interrogazioni.
4) Si calcoli la lunghezza di ricerca attesa supponendo che l'utente voglia 5 documenti rilevanti e che l'insieme dei documenti recuperati venga suddiviso nei seguenti 2 sottoinsiemi:
    - S1 contiene 5 documenti di cui 3 rilevanti e 2 non rilevanti
    - S2 contiene 5 documenti di cui 3 rilevanti e 2 non rilevanti
5) Si dia la definizione e il significato di precisione, richiamo e fallout per la misura dell'efficacia dei sistemi per il recupero delle informazioni.
6) Si descriva cosa è un thesaurus e quali caratteristiche possiede. Si descriva inoltre in particolare la relazione associativa e le sue diverse tipologie, riportando anche alcuni esempi.

Soluzione dell'esercizio 4)

L'utente legge i documenti in S1 e trova solo 3 documenti rilevanti sui 5 voluti, ma per farlo ha comunque dovuto esaminare anche gli altri 2 documenti nell'insieme, quindi: 3 documenti rilevanti trovati e 5 documenti letti finora.
Non avendo ancora trovato il numero di documenti rilevanti desiderati, deve leggere anche S2. A questo punto però gli basta trovare solo due documenti rilevanti dei 3 disponibili in S2, quindi il numero di documenti che l'utente deve esaminare in S2 dipende dalla posizione dei primi due documenti rilevanti nella lista dei 5 documenti in S2.
Non sapendo come il sistema di IR ordina i documenti all'interno dei sottoinsiemi, dobbiamo assumere che l'ordinamento in S2 sia casuale, e qui entra in gioco la teoria delle variabili casuali. Essendo l'ordinamento in S2 casuale, tutte le possibili combinazioni/ordini di tre documenti rilevanti e due non rilevanti hanno la stessa probabilità di essere fornite all'utente. Quindi si calcola il numero medio (o valore atteso) di documenti da leggere per trovare i primi due documenti rilevanti sui 5 facendo la media su tutti i possibili ordinamenti dei documenti in S2. Tutti i possibili ordinamenti sono 10:

1.	R	R	R	NR	NR	$\frac{3}{10}\times 2$
2.	R	R	NR	R	NR	
3.	R	R	NR	NR	R	
4.	R	NR	R	R	NR	$\frac{2}{10}\times 3$
5.	R	NR	R	NR	R	
6.	R	NR	NR	R	R	$\frac{1}{10}\times 4$
7.	NR	R	R	R	NR	$\frac{2}{10}\times 3$
8.	NR	R	R	NR	R	
9.	NR	R	NR	R	R	$\frac{2}{10}\times 4$
10.	NR	NR	R	R	R	

All'utente servono due soli documenti rilevanti quindi dobbiamo osservare la posizione dei primi due documenti rilevanti in ognuno degli ordinamenti e quanti documenti l'utente deve leggere per arrivare a tale posizione. Come si vede:

- nei primi 3 casi su 10 l'utente legge 2 soli documenti perché trova subito i documenti rilevanti;
- nei successivi 2 casi (il 4° e 5°) su 10 l'utente legge 3 documenti perché il secondo documento rilevante è al terzo posto;
- nel successivo caso (il 6°) su 10 l'utente legge 4 documenti;
- nei successivi 2 casi (il 7° e 8°) su 10 l'utente legge 3 documenti;
- nei successivi e ultimi 2 casi (il 9° e 10°) su 10 l'utente legge 4 documenti;

A questo punto, il valore atteso di documenti che l'utente deve leggere per trovare i primi due documenti rilevanti in S2 è la media del numero di documenti da leggere nei vari casi ma pesato per il numero di combinazioni per ognuno dei casi, cioè:

$$\frac{3}{10}\times 2+\frac{2}{10}\times 3+\frac{1}{10}\times 4+\frac{2}{10}\times 3+\frac{2}{10}\times 4=\frac{6}{10}+\frac{6}{10}+\frac{4}{10}+\frac{6}{10}+\frac{8}{10}=\frac{30}{10}=3$$

Questo, infine, va sommato al numero di documenti già letti dall'utente in S1, cioè:

$$5+3=8$$

La lunghezza di ricerca attesa è quindi 8.