Sistemi per il recupero delle informazioni
Laurea Magistrale in Editoria e Giornalismo
Prova scritta del 31 gennaio 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.
- Tempo a disposizione: 1 ora e 30 minuti.

1) Si descriva cosa è un sistema informativo, di cosa si occupa, come è composto e in che relazione è rispetto al sistema organizzativo e al sistema informatico.

2) Si citino i diversi tipi di interrogazione messi a disposizione dai sistemi per il recupero delle informazioni. Si descrivano in particolare le interrogazioni vettoriali, sottolineando i problemi che esse presentano, e dando brevemente l'idea su cui si basa il recupero dei documenti per questo tipo di interrogazioni.

3) Descrivere la legge di Zipf e cosa è una stop list.

4) Si calcoli la lunghezza di ricerca attesa supponendo che l'utente voglia 4 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 descrivano le misure di Richiamo e Precisione per valutare le prestazioni dei sistemi per il recupero delle informazioni e il rapporto che le lega.

6) (Facoltativo) Si descriva cosa è il processo di matching e come e perché si differenzia in matching e mapping. Si descriva poi in particolare il matching basato sulla prossimità.

Soluzione dell'esercizio 4)

L'utente legge i doc in S1 e trova solo 3 doc ril sui 4 voluti, ma per farlo ha comunque dovuto esaminare anche gli altri 2 doc nell'insieme, quindi: 3 doc ril trovati e 5 doc letti finora.

Non avendo ancora trovato il num di doc ril desiderati, deve leggere anche S2. A questo punto però gli basta trovare un solo doc ril dei 3 disponibili in S2, quindi il numero di documenti che l'utente deve esaminare in S2 dipende dalla posizione del primo doc rilevante nella lista dei 5 doc in S2.

Non sapendo come il sist. di IR ordina i doc 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 doc ril e due non rilevanti hanno la stessa probabilità di essere fornite all'utente. Quindi si calcola il numero medio (o valore atteso) di doc da leggere per trovare il primo doc ril sui 5 facendo la media su tutti i possibili ordini dei documenti in S2. Tutti i possibili ordini sono 10:

1.	R	R	R	NR	NR
2.	R	R	NR	R	NR
3.	R	R	NR	NR	R
4.	R	NR	R	R	NR
5.	R	NR	R	NR	R
6.	R	NR	NR	R	R
7.	NR	R	R	R	NR
8.	NR	R	R	NR	R
9.	NR	R	NR	R	R
10.	NR	NR	R	R	R

All'utente serve un solo doc ril quindi dobbiamo osservare la posizione del primo doc ril in ognuno degli ordinamenti e quanti doc l'utente deve leggere per arrivare a tale posizione. Come si vede:

- in 6 casi (i primi sei) su 10 l'utente legge 1 solo doc perché trova subito il doc ril;
- in 3 casi (dal 7° al 9°) su 10 l'utente legge 2 doc perché il doc ril è al secondo posto;
- in 1 caso (l'ultimo) su 10 l'utente legge 3 doc perché il doc ril si trova in terza posizione.

A questo punto, il valore atteso di doc che l'utente deve leggere per trovare il primo doc rilevante in S2 è la media del numero di doc da leggere nei tre casi ma pesato per il numero di combinazioni per ognuno dei casi, cioè:

$$\frac{6}{10}\times 1+\frac{3}{10}\times 2+\frac{1}{10}\times 3=\frac{6}{10}+\frac{6}{10}+\frac{3}{10}=\frac{15}{10}=1{,}5$$

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

$$5+1{,}5=6{,}5$$

La lunghezza di ricerca attesa è quindi 6,5.