INFORMATICA GENERALE – Esame del 19 febbraio 2014

COGNOME e NOME: _____MAT.:_____

In un motore di ricerca, ogni documento è solitamente descritto da un insieme di parole significative (*keyword*) presenti nel documento stesso. Ogni parola ha associato un peso che deriva dal conteggio opportunamente pesato della parola all'interno del documento. Tale descrizione consente di rappresentare il *corpus* dei documenti gestiti dal motore di ricerca in uno spazio vettoriale n-dimensionale dove n è il numero totale di keyword presenti nel corpus.

Ogni qual volta un utente interroga il motore di ricerca inserendo un insieme di keyword, la rilevanza di ogni documento presente nel corpus rispetto alla richiesta (*query*) viene così calcolata attraverso misure di similarità quali la similarità coseno (valore del coseno dell'angolo tra il vettore che rappresenta la query e il vettore che rappresenta ogni singolo documento).

Si vuole costruire un piccolo motore di ricerca in grado di gestire al più 20 documenti e 30 keyword di al più 100 caratteri l'una. Scopo del programma è restituire il documento più attinente alla query immessa dall'utente (L'algoritmo esposto è liberamente ispirato al modello vettoriale realmente funzionante in Google)

Testo problema	**Val.**	**Max**
Punto A: Scrivere la *funzione* `caricaKeyword` che permetta la lettura del file di testo *elencoparole.txt* e la memorizzazione nel vettore `char KW[30][100]` di *nCaratteristiche* parole; la funzione ritorna il numero di parole lette. Scrivere la *procedura* `caricaDocumenti` che legge il file di testo *elencocaratteristiche.txt* contenente la rappresentazione degli *nDocumenti* documenti presenti nel corpus quali vettori di *nCaratteristiche* pesi e lo memorizza nella matrice `float Doc[20][30]`. Ogni cella `Doc[i][j]`, per i=0…nDocumenti-1 e j=0…nCaratteristiche-1, contiene il peso della parola `KW[j]` nel documento i-esimo, se la parola è presente nel documento, 0, se la parola non è presente nel documento. Il `main`, dopo aver richiamato le funzioni di cui sopra dovrà presentare le opzioni (1) di stampa a video dell'elenco delle parole ricercabili contenute nel vettore `KW`, (2) di stampa a video delle rappresentazioni dei documenti nel corpus contenute nel vettore `Doc` e (5) di fine programma.		20
Punto B: Scrivere la procedura `costruisciQuery` che, dati come parametri di input una query, ovvero un vettore di keyword inserite dall'utente, la dimensione del vettore, e il vettore di tutte le keyword nel corpus `KW`, restituisca un vettore reale `Q[30]` di lunghezza reale *nCaratteristiche*, dove `Q[i]` contenente "1" se la parola `KW[i]` è presente nella query, "0" altrimenti. Modificare il main al punto (A) in modo che presenti l'opzione (3) che consenta ad un utente di inserire un insieme di parole chiave e, dopo aver richiamato la funzione `costruisciQuery`, stampi a video il vettore `Q`		4
Punto C: Scrivere la funzione `calcolaCoseno` che, dati come parametri formali due vettori di tipo reale e la loro dimensione effettiva, restituisca il coseno dell'angolo fra i due vettori (suggerimento: creare una funzione di supporto che calcoli il modulo di un vettore). Modificare il main al punto (A) in modo che presenti l'opzione (4) di calcolo del coseno fra il vettore `Q` ed un documento scelto dall'utente, e visualizzare tale risultato.		4
Punto D: Scrivere la funzione `trovaDocumento` che, dati come parametri formali la matrice `Doc` ed il vettore `Q`, ritorni il numero d'ordine del documento avente similarità coseno massima rispetto alla query. A tale scopo la funzione `trovaDocumento` dovrà richiamare la funzione `calcolaCoseno` per ogni documento nella matrice. Modificare il `main` al punto (A) in modo che modifichi l'opz (4) di stampa a video del numero del documento più rilevante rispetto alla query.		4
Voto		32

INFORMATICA GENERALE – Esame del 19 febbraio 2014

COGNOME e NOME: _____MAT.:_____

NOTE

La valutazione del codice prodotto avviene al termine della prova e punto per punto. Per ogni punto completato, è fondamentale che il codice compili e rispetti le specifiche descritte nel punto stesso.
Si ricorda che è possibile consultare solo il materiale cartaceo del corso e i libri di testo consigliati.

Suggerimenti

- Per evitare imprecisioni di calcolo, utilizzate sempre il tipo "double" anziché il tipo "float"
- La riga -- system("cls\n"); -- pulisce lo schermo (libreria: <stdlib.h>)
- Le ripetizione -- getchar();getchar(); -- permette di "ripulire" caratteri spuri ed attende una pressione sulla tastiera prima di proseguire l'azione (libreria: <stdlib.h>)
- "exit(0)" permette di interrompere l'esecuzione del programma (libreria: <stdlib.h>)
- "strcmp(stringa1,stringa2)" ritorna l'intero "0" se le due stringhe sono uguali (libreria: <string.h>)
- sqrt(*numero*) permette di avere la radice quadrata di *numero*; la libreria è: <math.h>

Definizioni utili

Modulo di un vettore: $$\left|\vec{X}\right| = \sqrt{\sum_{i=0}^{N-1} x_i^2}$$

Prodotto scalare di due vettori: $$X \cdot Y = \sum_{i=0}^{N-1} x_i y_i$$

Coseno nell'angolo fra due vettori: $$\cos(\vartheta) = \frac{X \cdot Y}{|X||Y|}$$

Esempio

inserendo le keyword "scienza tendenze rete"

- Punto B: il vettore Q è il seguente: [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1]
- Punto C: dando in ingresso il documento con identificatore 1 il valore del coseno è 0.090904
- Punto D:
  Il documento più vicino ha per identificatore il numero 9
  Il documento più vicino ha per coseno il valore 0.531585