Ingegneria del Software
Prova Scritta del 24 Luglio 2020

Tempo a disposizione: 20 minuti

Esercizio 1

Si consideri un software (**EasySimil**) che legge in ingresso più file contenenti testi diversi e li confronta con un altro file contenente un testo (*il file di esempio*), secondo una codifica per n-grammi. Un n-gramma è una sequenza di n caratteri consecutivi del testo. Se per esempio decidiamo n = 2 e che gli elementi sono caratteri un testo viene codificato contando quante volte compare ciascuna possibile coppia di caratteri. Ad esempio un testo potrebbe essere codificato con il seguente array:

aa	ab	ac	ad	ae	af	· · ·	zv	zw	zx	zy	zz
0	2	4	3	2	3	· · ·	0	0	0	1	3

Questa codifica consente di confrontare testi usando opportune formule di similarità. Ad esempio la seguente formula confronta il testo codificato dall'array T con il testo codificato dall'array S. L'operatore 'X' è un qualche operatore che opera su interi. Ad esempio si potrebbe usare l'operazione di sottrazione. Selezionando proprio questa operazione due testi sarebbero identici se (e solo se) il calcolo della similarità così computata sarebbe zero. Un valore grande indicherebbe invece dissimilarità.

$$\textbf{similarità} = \sum_{i=0}^{N-1} T[i] \times S[i]$$

Questo tipo di analisi, apparentemente elementare, cattura una sorprendente quantità di informazioni. Per esempio viene usata per identificare la lingua o addirittura distinguere l'autore di un testo. Ad esempio è possibile applicare **EasySimil** per confrontare sonetti di autori vari in lingue varie, ad esempio un sonetto di Dante con uno di Petrarca, uno di Shakespeare e uno di Goethe.

Selezionare uno stile architetturale tra quelli visti a lezione che sia adatto al problema sopra descritto e rappresentare con esso una possibile architettura software di **EasySimil.** Motivare la scelta dello stile fatta. Il sistema deve permettere all'utente di selezionare il numero 'n' usato per creare l'array. Inoltre il sistema deve permettere all'utente di selezionare diverse formule di similarità (un esempio è quella data sopra ma ne esistono altre). Infine il sistema deve gestire input e output (leggendo i vari file e stampando a video il file più simile rispetto a quello di esempio).