Sistemi informativi (corso progredito)
a.a. 2014/2015

Laboratorio n. 3

Reperimento dalla collezione sperimentale

Massimo Melucci

Obiettivi

- Comprensione dei meccanismi del reperimento.
- Un primo algoritmo di reperimento.

Base di partenza

- Le interrogazioni nel seguente formato:
  `Id.Interrogazione␣testo interrogazione`
- Le coppie nel seguente formato:
  `Id.Interrogazione␣parola interrogazione`
- Le coppie nel seguente formato:
  `Id.Interrogazione␣`*`stem`*` parola interrogazione`
- I giudizi di rilevanza.
- Il programma `trec_eval` utilizzato dal docente per misurare l'efficacia.
- I dati della lezione precedente.

Procedimento

Il procedimento è in un due passi, uno per ogni consegna:

- Alla prima consegna si deve produrre un documento di testo che illustra la propria funzione di reperimento, in uno dei modi seguenti:
  - Con un'espressione matematica del tipo, ad esempio,

$$f(x, y) = \sum_{i=1}^{k} x_i y_i$$

    dove x è, ad esempio, un vettore che rappresenta un documento e y è un vettore che rappresenta un'interrogazione. In questo modo, si devono fornire i dettagli necessari per poter implementare la funzione.
  - Con un algoritmo scritto in pseudo-codice in cui siano chiari i dettagli necessari per poter implementare l'algoritmo, cioè, quali sono le strutture di dati, i parametri e le eventuali funzioni di supporto necessarie. Si evitino, in questa fase, frammenti di codice scritto in un linguaggio di programmazione.

continua...

Procedimento

- ...
  - Con una descrizione in lingua italiana, ma con tutti i dettagli necessari per poter implementare la funzione. La descrizione deve essere il più rigorosa possibile e si deve controllare che ciò che si scrive sia poi effettivamente realizzabile.
  - Con una combinazione di due o più modi predetti.
  - In ogni caso, si verifichi che il tempo di esecuzione di un reperimento per una fissata interrogazione abbia un ordine di complessità del numero di descrittori dell'interrogazione o poco più; altrimenti, si giustifichi la maggiore complessità.
- Il nome del documento deve essere nel seguente formato: `lab-3-gruppo-`*n*`.txt`, dove *n* è il numero del gruppo; il formato può anche essere Microsoft Word (si usi l'estensione `.doc` o `.docx`) o Adobe PDF (si usi l'estensione `.pdf`).

Procedimento

- Alla seconda consegna, si devono produrre i seguenti risultati:
  - La versione definitiva del testo fornito alla prima consegna.
  - I risultati dell'esecuzione della funzione di reperimento nel seguente formato testuale:

    Id.Int. `Q0` Id.Doc. Rango Punteggio EtichettaRun

    dove Id.Int. è l'identificatore dell'interrogazione, `Q0` è costante, Id.Doc. è l'identificatore del documento, Rango è un numero naturale da 1 fino a 1000, Punteggio è un numero reale fornito dalla funzione di reperimento, EtichettaRun è l'etichetta che identifica la *run* nel formato `G`*n*`R`*m* dove *n* è il numero del proprio gruppo e *m* è il numero della run da 0 a 9; ad esempio `G17R3` è la *run* n. 3 del gruppo n. 17
  - **Attenzione**: non si aggiunga l'intestazione appena descritta.

continua...

Procedimento

- ...
- Ad esempio:

```
1   Q0   1234    1   1.234   G17R3
1   Q0   345     2   1.056   G17R3
1   Q0   2909    3   1.056   G17R3
...
1   Q0   12    114   0.034   G17R3
1   Q0   879   115   0.056   G17R3
1   Q0   3204  116   0.023   G17R3
2   Q0   12      1   3.467   G17R3
2   Q0   879     2   3.123   G17R3
...
```

- Si carichi la *run* in un file di testo omonimo, ad esempio, `G17R3.txt`