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

Laboratorio n. 2

Indicizzazione della collezione sperimentale

Massimo Melucci

Obiettivi

- Comprensione dei meccanismi dell'indicizzazione.
- Un algoritmo di indicizzazione.

Base di partenza

- La collezione di documenti tratti da numeri della rivista di informatica *Communications of the ACM* in formato XML e testo.
- Lista delle triple date da frequenza della parola nel documento, documento e la parola.
- Come sopra, ma con lo *stem* anziché la parola intera.
- Lista di *stopword*.
- Non è obbligatorio usare questi dati.

Procedimento

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

- Alla prima consegna si devono produrre:
  - un file con le parole chiave di ciascun documento nel seguente formato:
    `parola␣docid[␣peso1...␣pesoN]`
    dove `parola` è una parola chiave, `docid` è un identificatore di documento, ␣ indica uno spazio o un tabulatore, "[" indica l'inizio di una parte opzionale, "]" indica la fine di una parte opzionale, `pesoI` indica un peso numerico della parola nel documento; ad esempio, si può avere

`computer␣1234`	oppure `computer␣1234␣1`
`computer␣345`	`computer␣345␣4.5`
`network␣1234`	`network␣1234␣0.5`
`system␣4567`	`system␣4567␣-1.1`
  - un documento di testo che spiega brevemente come sono state scelte le parole e sono stati calcolati i pesi

continua...

Procedimento

...continua

- Alla seconda consegna si devono produrre:
  - un file simile al precedente, ma risultato da un processo automatico effettuato da strumenti *software* utilizzati a scopi di IR
  - un documento di testo che documenta brevemente l'algoritmo di indicizzazione che ha prodotto il file e il modo in cui si possono utilizzare gli strumenti per poter applicare l'algoritmo

  Per illustrare l'algoritmo è possibile utilizzare uno o una combinazione dei modi seguenti: lingua italiana o inglese, formule matematiche, pseudo-codice, codice R, codice Matlab; se possibile evitare codice scritto in C o Java.