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

Laboratorio n. 5

PageRank

Massimo Melucci

Obiettivi

- Comprensione di PageRank.
- Un'implementazione di PageRank.
- Una funzione di reperimento che integra PageRank.

Come funziona

Dopo aver implementato l'algoritmo e calcolato PageRank per ogni documento, ci sono diversi modi per integrare PageRank, ad esempio:

1. riordinamento:
   - 1.1 si produce il primo *ranking* (vedi laboratorio n. 2) con o senza RF, a scelta (vedi laboratorio n. 4)
   - 1.2 si fissa un numero $N \geq 1$
   - 1.3 si combina il punteggio della funzione di reperimento originale con PageRank e si calcola un nuovo punteggio, per ciascuno dei primi N
   - 1.4 gli altri documenti sono ordinati i primi N riordinati secondo il primo *ranking*
   - 1.5 si produce il secondo *ranking*
2. si inserisce PageRank nella funzione di reperimento in qualche modo "creativo", si applica la nuova funzione di reperimento e si produce un unico *ranking*

PageRank può essere integrato come probabilità stazionaria oppure come rango che ha un documento in una lista di documenti ordinata per probabilità.

Base di partenza

- Le citazioni tra documenti della collezione:
  - `citation.txt` nella forma documento-documento; si tenga presente esiste una coppia x, y se x cita y oppure se y cita x; inoltre, ci sono anche le coppie x, x; ad esempio

364	44
364	77
364	98
364	100
364	224
364	364

    vuol dire che 364 cita o è citato da `44 77 98 100 224 364`
  - `citation.list.txt` nella forma documento-lista-di-documenti; il file ha lo stesso contenuto dell'altro, ma in una forma diversa (ad esempio, `364 44 77 98 100 224 364` vuol dire che 364 cita o è citato da `44 77 98 100 224 364`).
- 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 integrata con PageRank. Il nome del documento deve essere nel seguente formato: `lab-5-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`).
- continua...

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 EtichettaRun è l'etichetta che identifica la *run* nel formato `G`*n*`R`*m*`PR` dove *n* è il numero del proprio gruppo, *m* è il numero della run da 0 a 9 e `PR` sta per PageRank; ad esempio `G17R3PR` è la *run* n. 3 del gruppo n. 17 con PageRank
- continua...

Procedimento

- ...
- Ad esempio:

```
1   Q0   1234    1   1.234   G17R3PR
1   Q0   345     2   1.056   G17R3PR
1   Q0   2909    3   1.056   G17R3PR
...
1   Q0   12    114   0.034   G17R3PR
1   Q0   879   115   0.056   G17R3PR
1   Q0   3204  116   0.023   G17R3PR
2   Q0   12      1   3.467   G17R3PR
2   Q0   879     2   3.123   G17R3PR
...
```

- **Attenzione**: non si aggiunga l'intestazione appena descritta.
- Si carichi la *run* in un file di testo omonimo, ad esempio, `G17R3PR.txt`