Come leggere la lista hackmeeting con TF-IDF

lucha@paranoici.org

24 giugno 2011

Leggendo un giorno la posta...

...un thread su lista hackmeeting

Tutti i tuoi movimenti su iphone?

DI COSA PARLA?

Leggendo un giorno la posta...

...un thread su lista hackmeeting

Tutti i tuoi movimenti su iphone?

DI COSA PARLA?

TF-IDF

Ci viene in aiuto un algoritmo: TF-IDF.

- *TF* term frequency.
- *IDF* inverse document frequency.

come capiamo di cosa stiamo parlando

riconosciamo una *parola chiave* associando un punteggo di *rilevanza* ad ogni termine.

TF-IDF

Ci viene in aiuto un algoritmo: TF-IDF.

- *TF* term frequency.
- *IDF* inverse document frequency.

come capiamo di cosa stiamo parlando

riconosciamo una *parola chiave* associando un punteggo di *rilevanza* ad ogni termine.

Term Frequency

Il primo fattore del punteggio è la term frequency:

$$\text{tf}(\text{parola}) = \frac{\text{occorrenze parola nel testo}}{\text{numero di parole nel testo}}.$$

Da sola *non* è rilevante: congiunzioni, avverbi, etc.

Inverse Document Frequency

Consideriamo una collezione di documenti che costituiscono un insieme di riferimento *generico* (ovvero che non individuano un argomento specifico che vogliamo distinguere).
Il secondo fattore del punteggio è la inverse document frequency:

$$\text{idf(parola)} = \frac{\text{numero totale di documenti}}{1 + \text{numero di documenti nei quali compare parola}}.$$

Il valore di idf non dipende dal testo che vogliamo analizzare, ma solo dalla collezione di documenti che consideriamo come base neutra. Può quindi essere precalcolato.

punteggio totale

tf-idf

$$\text{tf-idf} = \text{tf} * \text{idf} .$$

and the winner is