

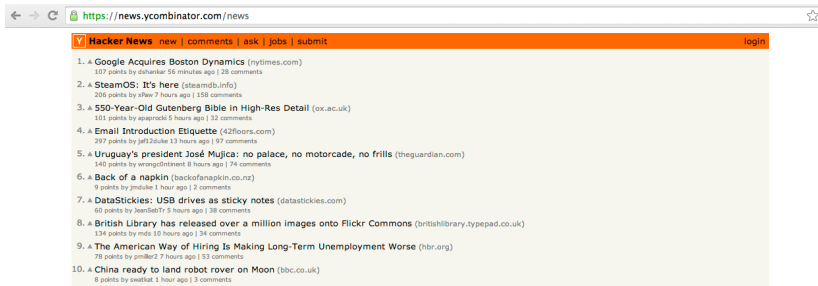
The Hacker News Similarity Distance

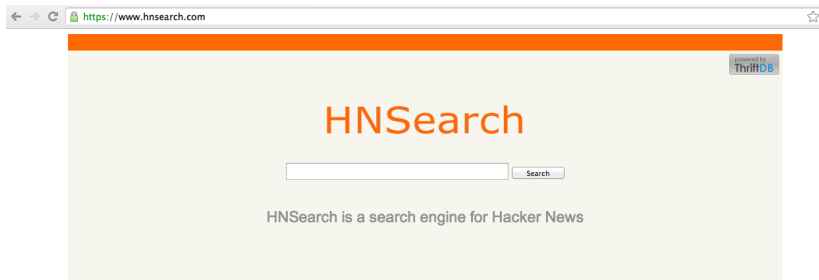
Jacopo Notarstefano

`jacopo.notarstefano [at] gmail.com`

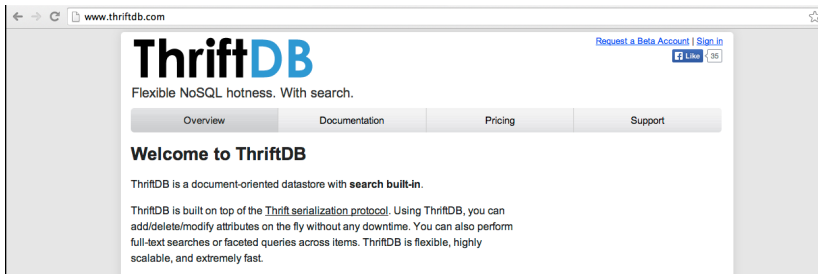
10 February 2014

Hacker News





ThriftDB



A simple RESTful API

The Google Similarity Distance

Definition (Cilibrasi, Vitányi 2007)

Let x and y be search terms and N the size of the Google index. We define:

$$\text{NGD}(x, y) = \frac{\max \{ \log f(x), \log f(y) \} - \log f(x, y)}{\log N - \min \{ \log f(x), \log f(y) \}}$$

where $f(x)$ denotes the number of pages containing x , $f(x, y)$ the number of pages containing both x and y as returned by Google.

The Hacker News Similarity Distance

Definition

Let x and y be search terms and N the size of the HNSearch index. We define:

$$\text{NHND}(x, y) = \frac{\max \{ \log f(x), \log f(y) \} - \log f(x, y)}{\log N - \min \{ \log f(x), \log f(y) \}}$$

where $f(x)$ denotes the number of pages containing x , $f(x, y)$ the number of pages containing both x and y as returned by HNSearch.

Implementation

Results