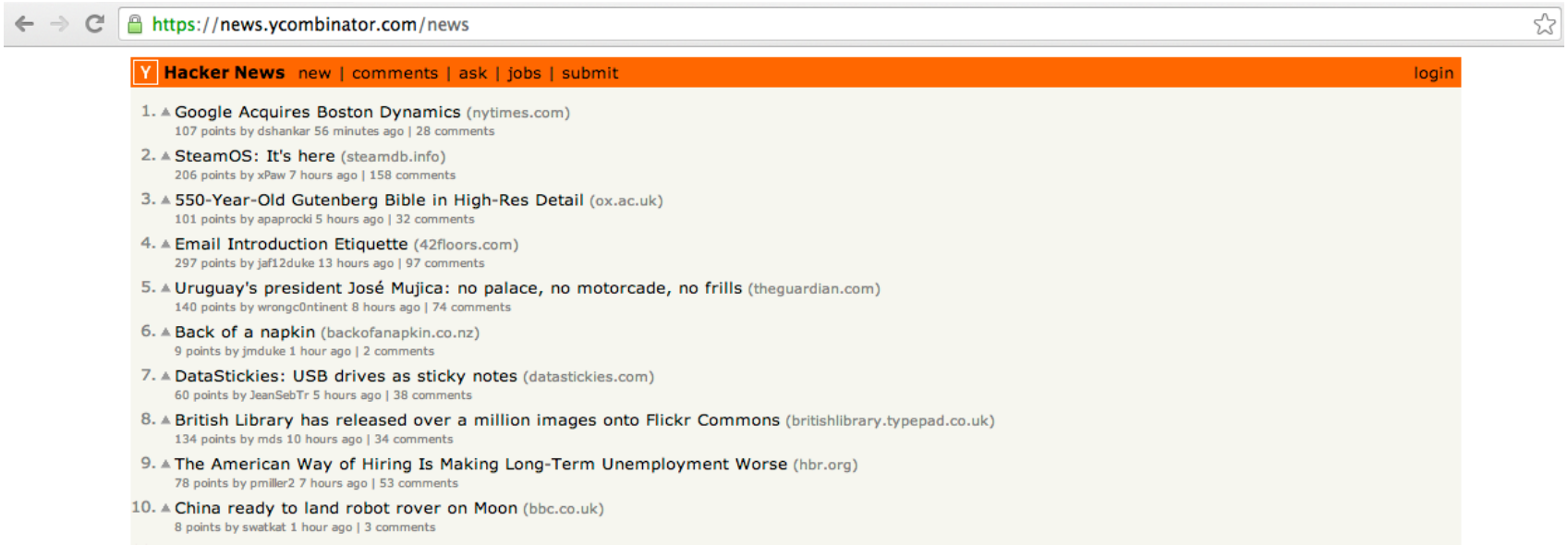
The Hacker News Similarity Distance

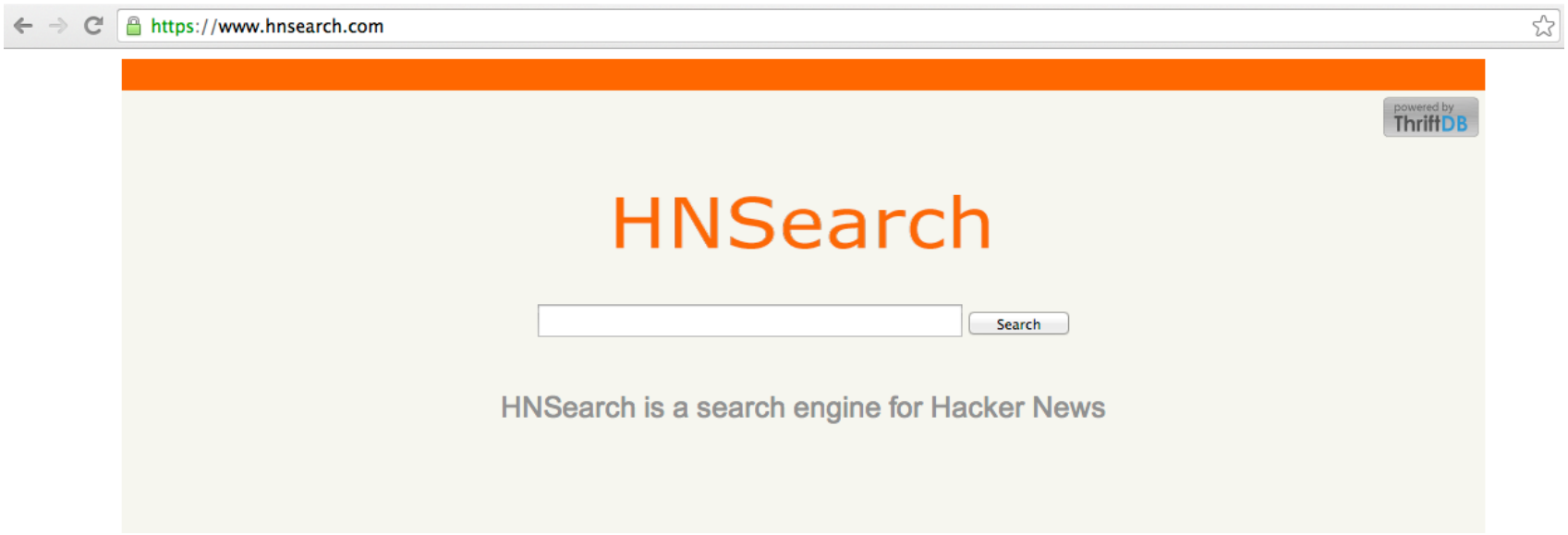
Jacopo Notarstefano
jacopo.notarstefano [at] gmail.com

10 February 2014

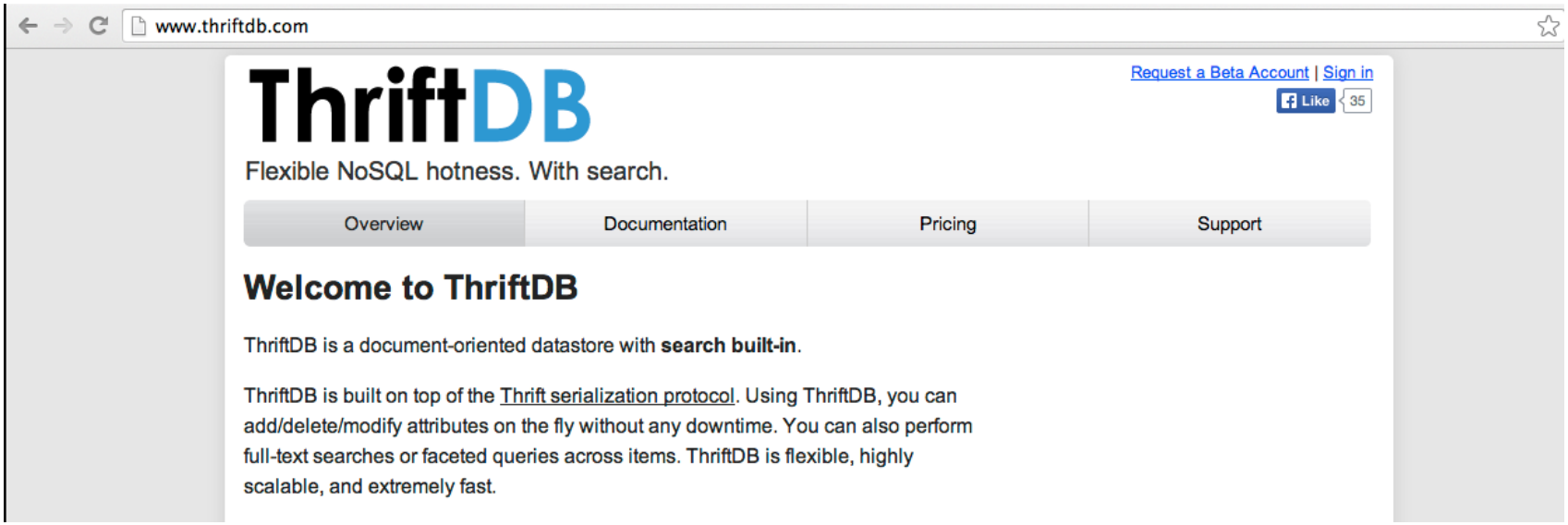
Hacker News


https://news.ycombinator.com/news
Hacker News new | comments | ask | jobs | submit login
1. Google Acquires Boston Dynamics (nytimes.com)
107 points by dshankar 56 minutes ago | 28 comments
2. SteamOS: It's here (steamdb.info)
206 points by xPaw 7 hours ago | 158 comments
3. 550-Year-Old Gutenberg Bible in High-Res Detail (ox.ac.uk)
101 points by apaprocki 5 hours ago | 32 comments
4. Email Introduction Etiquette (42floors.com)
297 points by jaf12duke 13 hours ago | 97 comments
5. Uruguay's president José Mujica: no palace, no motorcade, no frills (theguardian.com)
140 points by wrongc0ntinent 8 hours ago | 74 comments
6. Back of a napkin (backofanapkin.co.nz)
9 points by jmduke 1 hour ago | 2 comments
7. DataStickies: USB drives as sticky notes (datastickies.com)
60 points by JeanSebTr 5 hours ago | 38 comments
8. British Library has released over a million images onto Flickr Commons (britishlibrary.typepad.co.uk)
134 points by mds 10 hours ago | 34 comments
9. The American Way of Hiring Is Making Long-Term Unemployment Worse (hbr.org)
78 points by pmiller2 7 hours ago | 53 comments
10. China ready to land robot rover on Moon (bbc.co.uk)
8 points by swatkat 1 hour ago | 3 comments

HNSearch


https://www.hnsearch.com
powered by ThriftDB
HNSearch
Search
HNSearch is a search engine for Hacker News

ThriftDB


www.thriftdb.com
ThriftDB
Request a Beta Account | Sign in
Like 35
Flexible NoSQL hotness. With search.
Overview
Documentation
Pricing
Support
Welcome to ThriftDB
ThriftDB is a document-oriented datastore with search built-in.
ThriftDB is built on top of the Thrift serialization protocol. Using ThriftDB, you can add/delete/modify attributes on the fly without any downtime. You can also perform full-text searches or faceted queries across items. ThriftDB is flexible, highly scalable, and extremely fast.

A simple 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:

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

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:

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

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