HACKER NEWS SEARCH ENGINE

CS6200 Final Project

Suhas Mohan

Abstract

Hacker News is a social news website, similar to Reddit, focusing on computer science and entrepreneurship. It is run by Paul Graham's investment fund and startup incubator, Y Combinator. However, unlike Reddit where new users can immediately both up-vote and down-vote content, Hacker News does not allow users to down-vote content until they have accumulated 501 "karma" points. Karma points are calculated as the number of upvotes a given user's content has received minus the number of downvotes. "Flagging" comments, likewise, is not permitted until a user has 30 karma points.
Users can post Stories (Links or Text), comments and can upvote/downvote stories or comments.

Current Solutions

Hacker News does not have a default search engine. Currently the Hacker News search is powered by Algolia, a 3rd party search provider. This allows you to search for posts by title, text in comments or both. The default search option is set to “Stories” and the user has to explicitly select “All” in order to search on both stories and comments.
There is also no snippet generation so the entire comment/story text is displayed on the search engine result page.

My Hacker News search engine will use the official Hacker News dataset and present users with a clean and simple UI that will allow them to search for both stories and comments without explicitly choosing between the two.

Dataset

The official Hacker News data set is available on Kaggle and the Google Cloud Big Query platform. It can be found at https://www.kaggle.com/hacker-news/hacker-news.
The dataset contains all stories and comments from Hacker News from its launch in 2006. Each story contains a story id, the author that made the post, when it was written, and the number of points the story received.

The dataset has 2 main tables – Stories and Comments
The stories table is 402MB and has ~2Mil rows
The Comments table is 3.41GB and has ~8.4Mil rows

Stories

Contains information about all stories posted on Hacker News

Field name	Type	Mode	Description
id	INTEGER	NULLABLE	Unique story ID
by	STRING	NULLABLE	Username of submitter
score	INTEGER	NULLABLE	Story score
time	INTEGER	NULLABLE	Unix time
time_ts	TIMESTAMP	NULLABLE	Human readable time in UTC (format: YYYY-MM-DD hh:mm:ss)
title	STRING	NULLABLE	Story title
url	STRING	NULLABLE	Story url
text	STRING	NULLABLE	Story text
deleted	BOOLEAN	NULLABLE	Is deleted?
dead	BOOLEAN	NULLABLE	Is dead?
descendants	INTEGER	NULLABLE	Number of story descendants
author	STRING	NULLABLE	Username of author

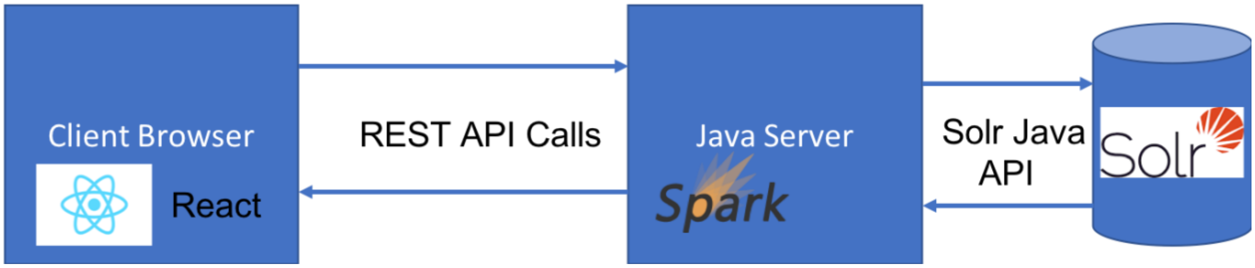
Comments

Contains information about all the comments on each story.

Field name	Type	Mode	Description
id	INTEGER	NULLABLE	Unique comment ID
by	STRING	NULLABLE	Username of commenter
author	STRING	NULLABLE	Username of author
time	INTEGER	NULLABLE	Unix time
time_ts	TIMESTAMP	NULLABLE	Human readable time in UTC (format: YYYY-MM-DD hh:mm:ss)
text	STRING	NULLABLE	Comment text
parent	INTEGER	NULLABLE	Parent comment ID
deleted	BOOLEAN	NULLABLE	Is deleted?

Field name	Type	Mode	Description
dead	BOOLEAN	NULLABLE	Is dead?
ranking	INTEGER	NULLABLE	Comment ranking

Architecture



All documents are indexed in Solr using a custom schema and results are retrieved using a custom ranking function.

The search engine is developed as a Web Application.

The front end is written using the ReactJS and Bootstrap frameworks. The front end communicates with the server using HTTP REST API calls.

The Java REST server is developed using the Spark Framework. It reads the requests from the client, creates a query for Solr, fetches results and creates a JSON response to send back to the client.

Custom Solr Schema

The Hacker News data is indexed in Solr using a custom schema to facilitate better indexing and support a custom ranking function.

Field Name	Type
ID	string
Title	text_general
Text	text_general
Author	text_general
HN_Score	Int
Type	String
url	string

ID – Unique comment/story ID given by Hacker News
Title – The title of the story
Text – The text of the story/comment
Author – The username of the Author for the comment/story
HN_Score – Score of the comment/story. Determined by the number of upvotes and downvoted made on the post.
Type – Comment/Story
URL – The URL of the story (if present)

Custom Ranking Function

A Custom ranking function is used to provide end users with more relevant results.
Solr's default BM25 ranker is augmented with data from Hacker News.
More weight is given to results that have the search terms present in the title of the post. The results are further boosted using the HN_Score field in Solr.

The HN_Score is given by Hacker News and is calculated using the number of upvotes and downvotes on a story/comment. A user is not allowed to downvote a post until they have accumulated 501 "karma points". Karma points are calculated as the number of upvotes a given user's content has received minus the number of downvotes.
Therefore, the Hacker News Score is a good indicator of the quality of the post and posts that have a high score are likely more relevant to the end user.

The final post score is calculated as:

$$\text{Score}(d_i) = \text{BM25}(d_i) * (\text{Title}^{\wedge}1.5 + \text{Text}^{\wedge}1.2 + \text{HN\_Score})$$

Features Implemented

Apart from providing users with search results for their queries, several other features have been implemented.

Snippet Generation

Relevant snippets that contain the search term are displayed in every search result. The snippets are up to 200 characters in length.

Search Term Highlighting

The user’s search terms are highlighted in the snippet and title of the post. This helps the user quickly identify what they are searching for and find the most relevant link to click on.

Pagination of Results

Results are divided into pages of 10 results each. The user can click on “previous”, ”next” to view then previous or next 10 results.

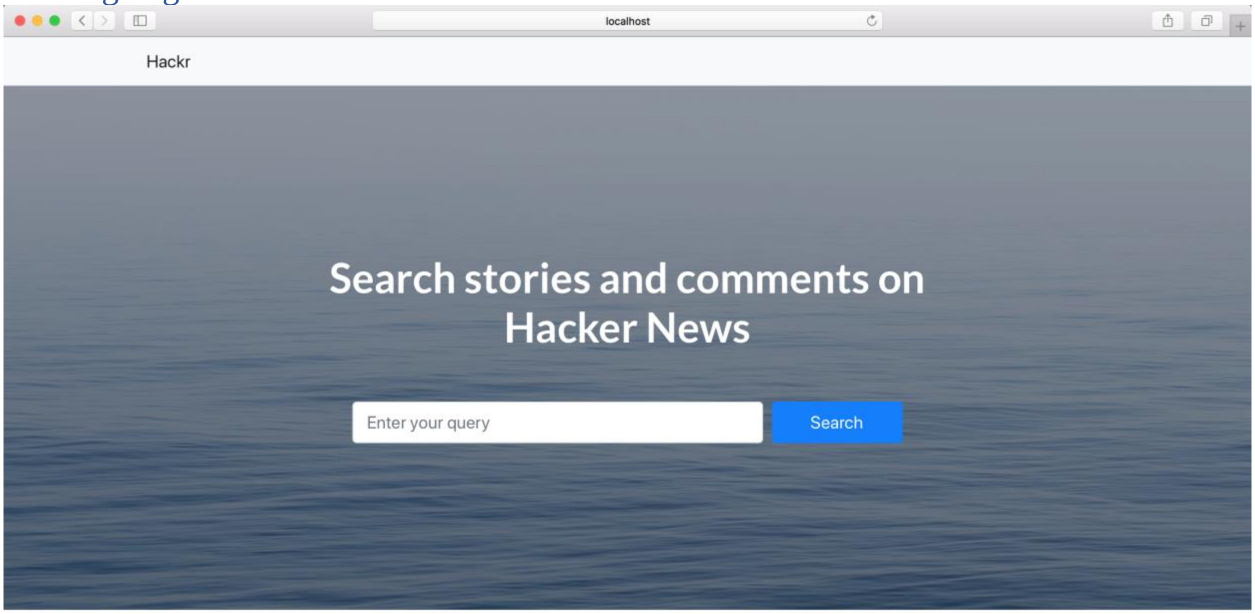
Sorting of Results

The user can choose to sort the search results using one of 3 options:

- Relevance – Default sort option. This is the order returned by the custom ranking function
- Date – Sort by newest story/comment first
- Score – Sort by highest Hacker News score first

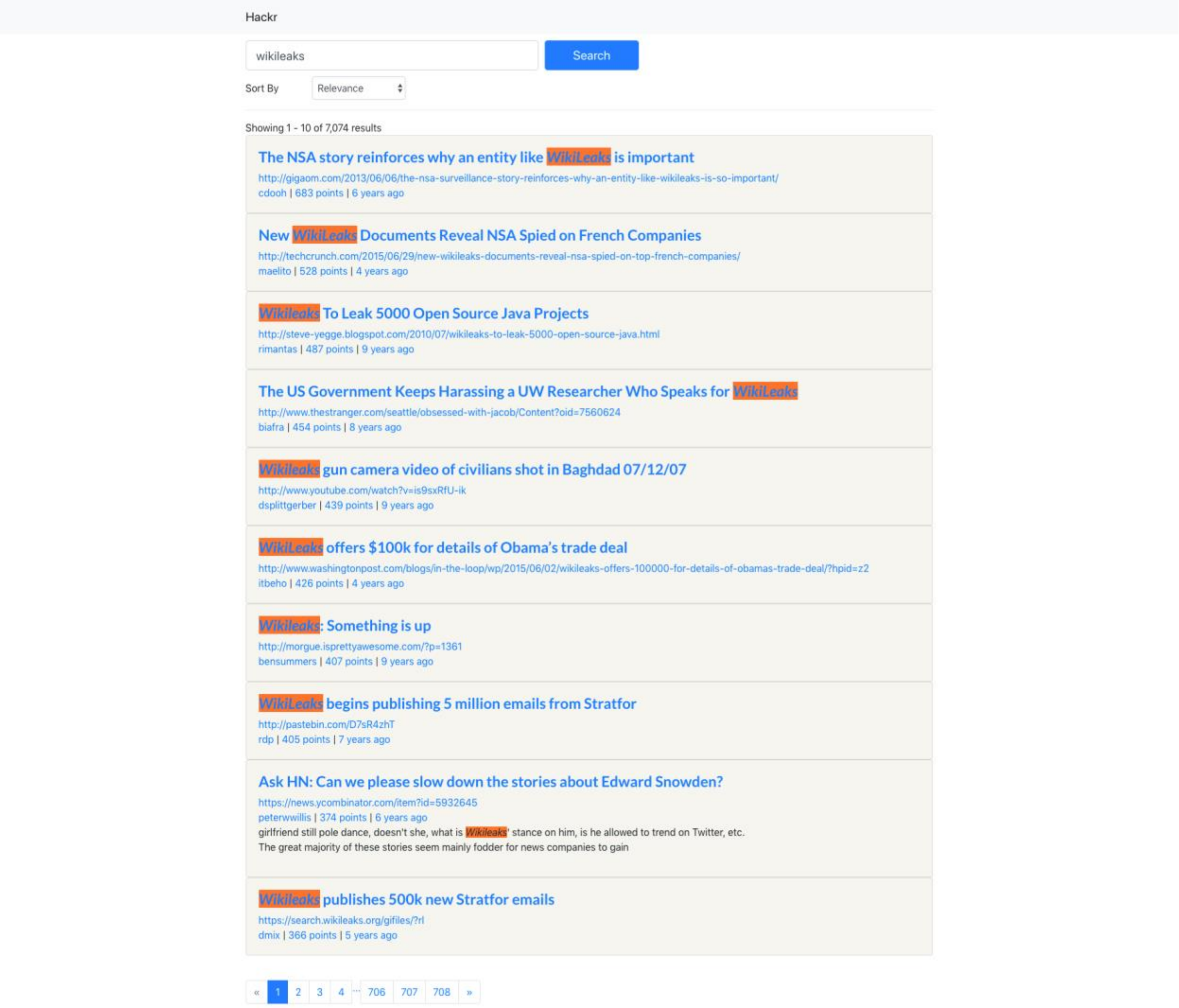
Screenshots

Landing Page



Search stories and comments on Hacker News

Search Engine Results Page



wikileaks

Sort By Relevance

Showing 1 - 10 of 7,074 results

The NSA story reinforces why an entity like WikiLeaks is important
http://gigaom.com/2013/06/06/the-nsa-surveillance-story-reinforces-why-an-entity-like-wikileaks-is-so-important/
cdooh | 683 points | 6 years ago

New WikiLeaks Documents Reveal NSA Spied on French Companies
http://techcrunch.com/2015/06/29/new-wikileaks-documents-reveal-nsa-spied-on-top-french-companies/
maelito | 528 points | 4 years ago

Wikileaks To Leak 5000 Open Source Java Projects
http://steve-yegge.blogspot.com/2010/07/wikileaks-to-leak-5000-open-source-java.html
rimantas | 487 points | 9 years ago

The US Government Keeps Harassing a UW Researcher Who Speaks for WikiLeaks
http://www.thestranger.com/seattle/obsessed-with-jacob/Content?oid=7560624
biafra | 454 points | 8 years ago

Wikileaks gun camera video of civilians shot in Baghdad 07/12/07
http://www.youtube.com/watch?v=is9sxRfU-ik
dsplittgerber | 439 points | 9 years ago

WikiLeaks offers $100k for details of Obama's trade deal
http://www.washingtonpost.com/blogs/in-the-loop/wp/2015/06/02/wikileaks-offers-100000-for-details-of-obamas-trade-deal/?hpid=z2
itbeho | 426 points | 4 years ago

Wikileaks: Something is up
http://morgue.isprettyawesome.com/?p=1361
bensummers | 407 points | 9 years ago

WikiLeaks begins publishing 5 million emails from Stratfor
http://pastebin.com/D7sR4zhT
rdp | 405 points | 7 years ago

Ask HN: Can we please slow down the stories about Edward Snowden?
https://news.ycombinator.com/item?id=5932645
peterwwillis | 374 points | 6 years ago
girlfriend still pole dance, doesn't she, what is Wikileaks' stance on him, is he allowed to trend on Twitter, etc.
The great majority of these stories seem mainly fodder for news companies to gain

Wikileaks publishes 500k new Stratfor emails
https://search.wikileaks.org/gifiles/?rl
dmix | 366 points | 5 years ago