TECHNISCHE UNIVERSITÄT BERLIN

AIM-3: SCALABLE DATA ANALYSIS & DATA MINING

Unsupervised Clustering of Hacker News Stories

Authors:
Bram LEENDERS & Marc ROMEYN

June 21, 2015

1 Introduction

1.1 What is HackerNews

Hacker News is a social news site: it aggregates news by allowing users to submit stories. Interesting submissions can be upvoted by other users and all submissions are ranked by popularity.
The content on Hacker News is mostly related to science, in particular computer science. The guidelines for what content can be posted are very broad; the guidelines specify on topic as "*anything that gratifies one's intellectual curiosity*"[1].
Over the last seven years Hacker News has experienced rapid growth, resulting in a daily 2.6 million pageviews and 3.5 unique visitors per month[2]. One of the reasons suggested for this popularity is Hacker News' similarity to how Reddis used to be[3]: user-submitted content with a very minimalistic, terminal-like interface.

1.2 Research Question

Run two clustering algorithms, compare results, profit.
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2 Dataset

As explained in the previous section, our analysis is on the set of all stories posted on Hacker News.
Let us first consider what we mean when we refer to a story. To do so, we have to separate between the various types of content on Hacker News, which can be catagorized in these four groups:

- Story: the majority of submissions are stories. A story can be either a link to another webpage or a short text by the submittor.

[1] `https://news.ycombinator.com/newsguidelines.html`
[2] `https://news.ycombinator.com/item?id=9219581`
[3] `http://techcrunch.com/2013/05/18/the-evolution-of-hacker-news/`

- Job: companies sponsored by YCombinator (the seed investor behind Hacker News) can post job offers on Hacker News. The percentage of Jobs is very small: well under one percent of the total volume.
- Poll: users can submit multiple choice questions for other users to answer.
- Comment: the three types mentioned above can receive comments by other users.

Since this research focuses only on the stories, we have left out the other three types. The dataset used in our research contains all stories between February 19th 2007 (the date Hacker News was launched[4]) and June 10th 2015 (the day we ran our crawler). This is a time span of 3033 days, during which a total of over 1.5 million stories were submitted.

2.1 Data Collection

To crawl all stories, we used the official Hacker News API[5]. This API returns some basic data about the story, such as the submitter, title, points (upvotes), a (possibly empty) story text and a (possibly empty) url. Stories generally have either a story text or a url, most only have a url.
For the stories with a non-empty story text field, fetching it from the API is enough. For other stories, we also fetched the content the URL links to. An important remark here is that some urls (especially the old ones) have become invalid over the course of time.
Something about the article extractor GoOse
We split the result in files with 1 day of articles per file. This dataset is about 4.4 GB in size.

3 Initial Analysis

We have now established some semantics of the data gathered for our research. Before diving into the data analysis, we will first quantify the data by providing the reader with some statistics.

3.1 Overall Activity

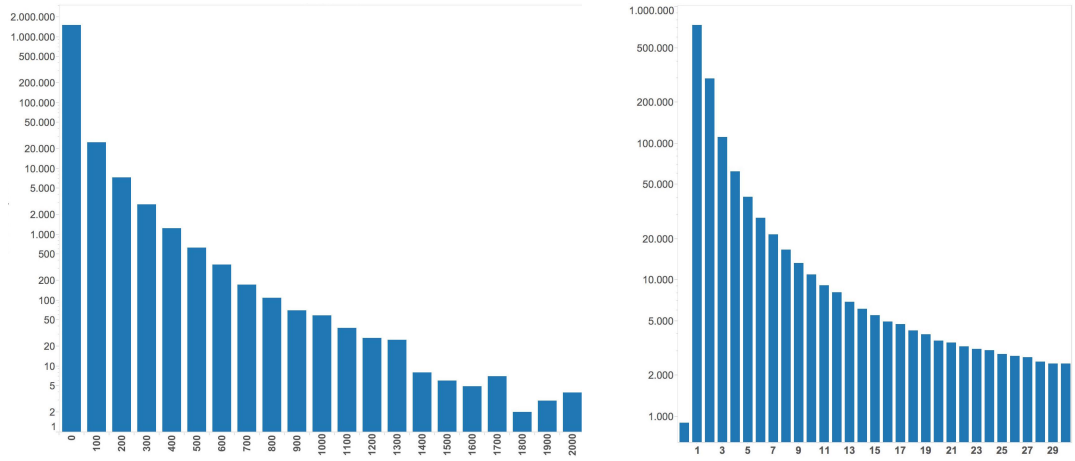
Statistics	
Stories	1544261
Submitters	165126
Upvotes	16668848
Comments	7383865

An interesting remark, is that January 6th 2014 did not have any submissions. The reason for this was a long downtime of the Hacker News website [6].

[4] `https://news.ycombinator.com/hackernews.html`
[5] `https://hn.algolia.com/api`
[6] `https://twitter.com/hackernewsonion/status/420068968464789505` - Hacker News is DOWN, but your chances of getting into YC if you know how to scale a plain text website are UP.

3.2 Upvote Distribution

If users like a story, they can express their approval by upvoting it. These upvotes are then used to rank the stories by popularity and calculate the (currently) most popular submissions. Since the formula for determining a posts "score" strongly favours newly submitted stories, even stories with few upvotes can spike to the frontpage.
To enter the frontpage (the top 30), a story often needs at least ten upvotes, stories that stay on the frontpage for a longer period of time will often have over a hundred upvotes. The distribution of upvotes per story can be seen in figure 3.1. Stories placed in buckets of size 100 based on the number of upvotes they received. The graph plots the size of each bucket
Note that these graphs are scaled logarithmically: in the graph showing all posts (figure 3.1a), 97.5% of the stories are in the first bucket. That means only 2.5% of the stories get over 100 upvotes.
The second graph (3.1b) is a zoomed in version with a bucket size of 1, and shows the distribution for points with 30 upvotes or less, a set of over 1.4 million stories.



(a) Distribution over all stories
Bucket size 100

(b) Distribution over low-ranked stories
Bucket size 1

Figure 3.1: Distributions of upvotes over posts (log scale)

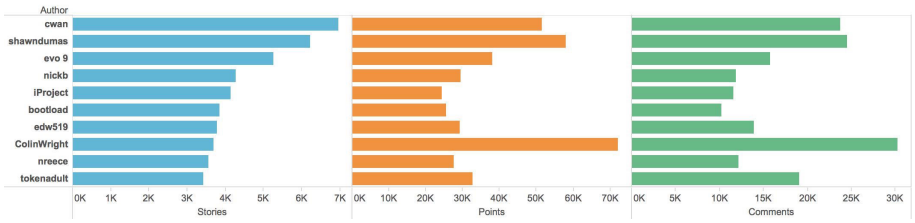
There are only six stories with over 2100 upvotes. For completeness' sake, these are the stories whose popularity is quite literally "off the charts":

Most upvoted stories	
Title	Points
Steve Jobs has passed away	4271
Tim Cook Speaks Up	3086
2048	2732
Don't Fly During Ramadan	2617
Hyperloop	2549
Microsoft takes NET open source and cross platform	2376

3.3 Top Users

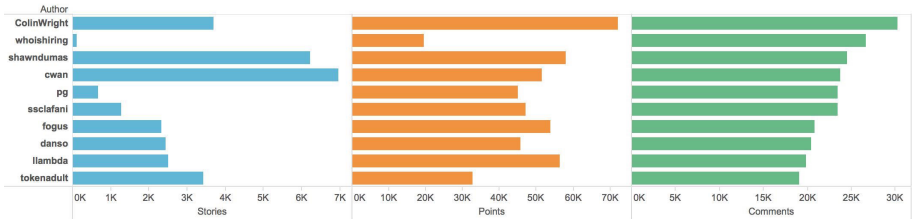
In figure 3.2, we show the statistics for some of the most active submitters. These statistics illustrate that Hacker News has a very active group of users. Together, this top 10 submitted 45,190 stories, which equals an average of 1.5 stories per user per day over the last 7 years.

Figure 3.2: Top 10 submitters by submitted stories



If we take the top 10 not by the number of stories a user has submitted but by the number of comments his or her stories have received, we get the rankings in figure 3.3.

Figure 3.3: Top 10 submitters by received comments



The interesting second place in this top 10 is user "_whoishiring". This user has almost no posts (a mere 114) but received 26,649 comments. Experienced Hacker News readers may have already expected this, for this user starts a monthly thread where companies can post job offers and others can respond to these. As such, the user only posts one story per month but its stories are discussed very actively.

3.4 Top Domains

As already stated in the description of the dataset, many of the stories are links to external sites. To provide some insights into which sites attract a lot of attention from the Hacker News community, we made three top ten lists that rank the sites by the same criteria as we did in the previous section: by number of stories, by points and by number of comments.

Most popular domains		
By Number of Stories	By Points	By Comments
techcrunch.com (27.711)	github.com (400.373)	techcrunch.com (172.854)
github.com (26.596)	techcrunch.com (365.207)	nytimes.com (153.471)
youtube.com (21.977)	nytimes.com (287.234)	github.com (127.812)
nytimes.com (18.125)	arstechnica.com (176.633)	arstechnica.com (80.666)
medium.com (14.172)	wired.com (161.698)	wired.com (75.406)
arstechnica.com (12.657)	medium.com (132.468)	washingtonpost.com (56.257)
wired.com (10.867)	bbc.co.uk (98.031)	medium.com (53.924)
bbc.co.uk (8.118)	washingtonpost.com (97.537)	bbc.co.uk (51.828)
en.wikipedia.org (7.058)	youtube.com (96.339)	theatlantic.com (41.530)
businessinsider.com (6.877)	theatlantic.com (77.628)	online.wsj.com (36.729)

These rankings provide some insights into what types on news are popular. The big geeky news sites (Techcrunch, Ars Technica and Wired) are present and are about as popular as the big newspapers (NY Times, BBC, Washington Post).
The high ranking of github.com (a code hosting site, *not* a news site) can be explained by the large number of open source projects hosted on GitHub that submit links to new versions and press releases on Hacker News. Some examples of these projects are Facebook's React framework, Twitter's Bootstrap, SQLite and io.js.
Notably, GitHub's competitors (e.g. BitBucket, GitLab and Beanstalk) are not in these rankings. This demonstrates GitHub's overpowering popularity in the code hosting market, at least in the open source community.

4 Analysis

Say that we did two types of unsupervised clustering: k-means and LDA. K-means

4.1 Latent Dirichlet Allocation

Some explanation what the F this is

4.2 Implementation

5 Results

$1 + 1 = 10$

6 Conclusion

We‘re awesome!