



L-Università  
ta' Malta

# Analyzing Topic Prevalence in Popular Hacker News Stories

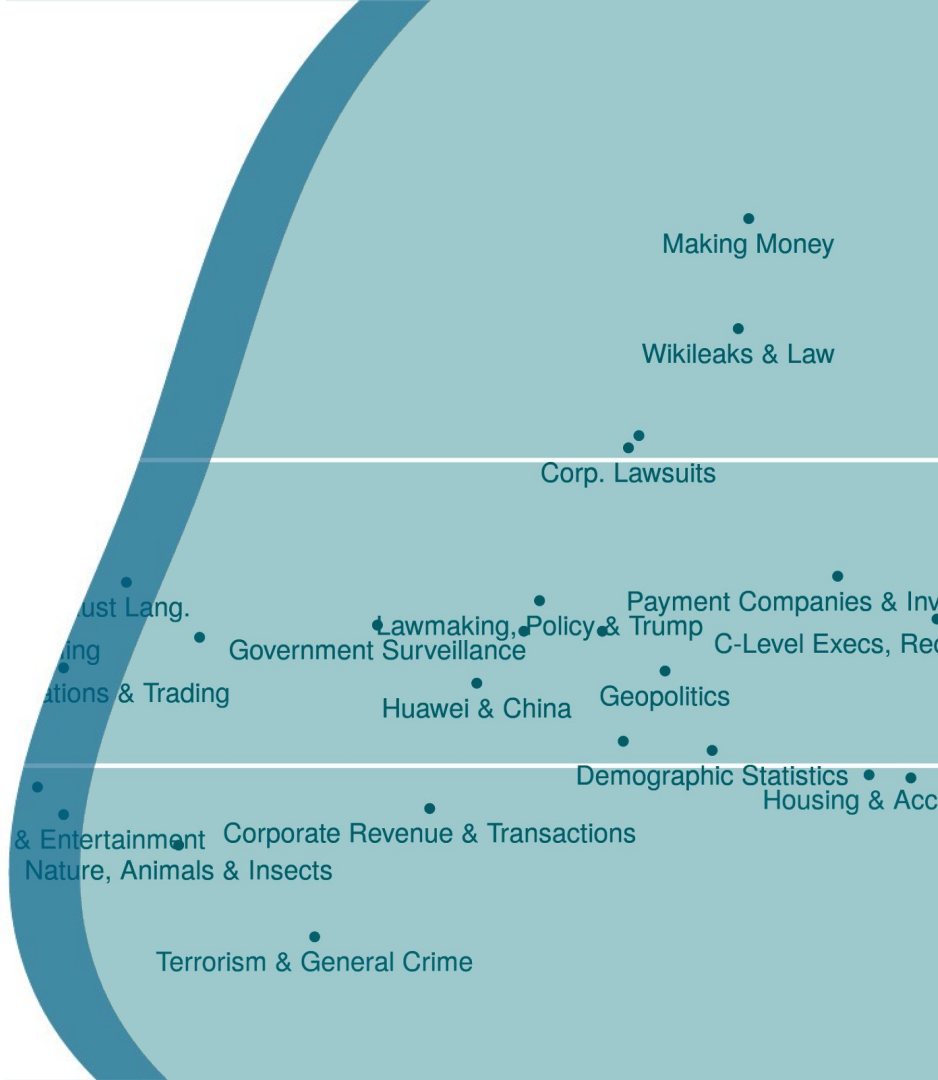
ICS5115 - Statistics for Data Scientists

Jake J. Dalli

[jake.dalli.10@um.edu.mt](mailto:jake.dalli.10@um.edu.mt)

MSc. Artificial intelligence

Faculty of ICT



## Data Collection

- We perform a one-time extract from the HN archive HckrNews, which includes the top 50% articles by points between **1st December 2018** and **31st May 2019**
- The extract is parsed to create a CSV file containing all information on the bulletin board (headline, link, points, comments, date, domain)
- The top 30 posts by points by day are filtered, associating a rank number to the post
- We iteratively scrape the page for each source and store the source as an HTML document within a folder using the Rvest R Library
- *Output dataset: hackernews\_board.csv*

	Hacker News	new	past	comments	ask	show	jobs	submit
1.	▲	Open-sourcing Sorbet: a fast, powerful type checker for Ruby (sorbet.org)	284 points by amandeev 3 hours ago   hide   64 comments					
2.	▲	Google Maps is filled with false business addresses pretending to be nearby (11 points by larsp 2 hours ago   hide   55 comments)						
3.	▲	Unseen 9/11 photos bought at house clearance sale (bbc.com)	59 points by yitchele 20 minutes ago   hide   1 comment					
4.	▲	Tokyo's suburban housing became vast ghettos for the old (theguardian.com)	23 points by nevertheless 58 minutes ago   hide   10 comments					
5.	▲	Song of the Rarest Large Whale on Earth Recorded for the First Time (gizmodo.com)	51 points by autabill 2 hours ago   hide   23 comments					
6.	▲	ECMO pumps blood out of the body, oxygenates it and returns it to the body (37 points by howard041 2 hours ago   hide   21 comments)						
7.	▲	Firefox zero-day was used in attack against Coinbase employees, not its user (125 points by qd-rn 6 hours ago   hide   43 comments)						
8.	▲	Support for right-to-repair laws slowly grows (arstechnica.com)	148 points by jaredperkins 8 hours ago   hide   108 comments					
9.	▲	Gryphon: An open-source framework for algorithmic trading in cryptocurrency (91 points by rrr 2 hours ago   hide   26 comments)						
10.	▲	Forget monoliths vs. microservices: cognitive load is what matters (techbeacon)	182 points by fercyfish 4 hours ago   hide   29 comments					
11.	▲	Personal data of 2.9M people leaked from Desjardins (cbc.ca)	55 points by timbaite 2 hours ago   hide   10 comments					
12.	▲	Slack Is Going Public Without an IPO – How a Direct Listing Works (fortune.com)	43 points by sransethi 5 hours ago   hide   3 comments					
13.	▲	In Japan, It's a Riveting TV Plot: Can a Worker Go Home on Time? (nytimes.com)	110 points by pseudolus 8 hours ago   hide   103 comments					
14.	▲	Civic honesty around the globe (sciencemag.org)	13 points by spivna 2 hours ago   hide   52 comments					
15.	▲	Why Monzo's bank transfers weren't working on the 30th of May (monzo.com)	126 points by rmlm 11 hours ago   hide   18 comments					
16.	▲	Sony launches a taxi-hailing app in Tokyo (techcrunch.com)	124 points by jay-paul 11 hours ago   hide   103 comments					
17.	▲	Capt. "Sully" Sullenberger Slams Boeing for Inadequate Pilot Training (yahoo.com)	51 points by jae21 1 hour ago   hide   6 comments					
18.	▲	The murky business of crystal mining (theguardian.com)	4 points by laurus 35 minutes ago   hide   discuss					
19.	▲	Flexport is hiring software engineers in Chicago and San Francisco (flexport.com)	1 hour ago   hide					
20.	▲	Bronx High School Math Bulletin (1957) [pdf] (lampart.azurewebsites.net)	14 points by amyltix 3 hours ago   hide   9 comments					
21.	▲	The Hemingway Marlin Fish Tournament (theparisreview.org)	0 points by danielw 2 hours ago   hide   discuss					
22.	▲	Oscar Wilde's talk inspired his rise and led to his downfall (instimes.com)	20 points by stromberg 5 hours ago   hide   10 comments					
23.	▲	Tesla Model 3 spoofed off the highway (regulus.com)	150 points by stromberg 5 hours ago   hide   110 comments					
24.	▲	A Hundred Unchildlike Lullabies (bookforum.com)	26 points by hermostov 4 hours ago   hide   2 comments					
25.	▲	BonziBuddy (wikipedia.org)	74 points by amertf 4 hours ago   hide   46 comments					
26.	▲	Getting 2FA Right in 2019 (hackernews.com)						

## Methodology & Key Assumptions

- **Descriptive Analysis:** Visual & Correlation
  - What is the relationship between points and comments?
  - Do temporal factors affect ranking?
- **Topic Analysis:** Detect Topics from the Corpus
  - Using a *Bayesian Generative Model* to detect topics
  - We find out which topics occur most frequently together
  - Which topics occur more in the top rank than other ranks
- We assume that the top page is adequately represented by the top 30 articles by points from the archive, and also disregard videos, pdfs and obscure web applications

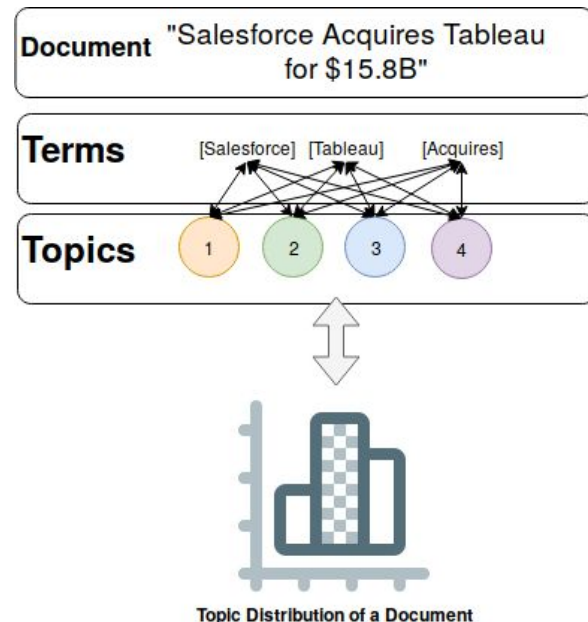
Data Sources
<a href="https://hckrnews.com/">https://hckrnews.com/</a>
Several News Websites

Source Datasets
hackernews_board.csv
hackernews_board_content.csv

## Latent Dirichlet Allocation

- Represents documents as mixtures of latent topics by term distributions
- Assumes prior knowledge of  $K$  (number of topics to detect).
- Therefore tightly coupled with a *parameter estimation problem*
- We evaluate the mixture by calculating the *Maximum Log Likelihood*

$$L(W) = \log P(d|\Upsilon, \alpha) = \sum_t \log P(d_t|\Upsilon, \alpha)$$

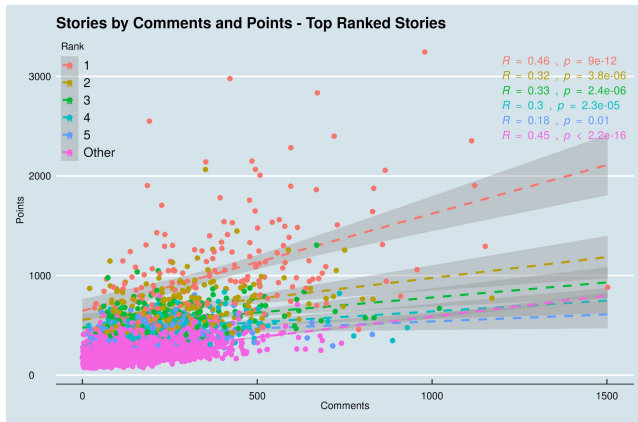
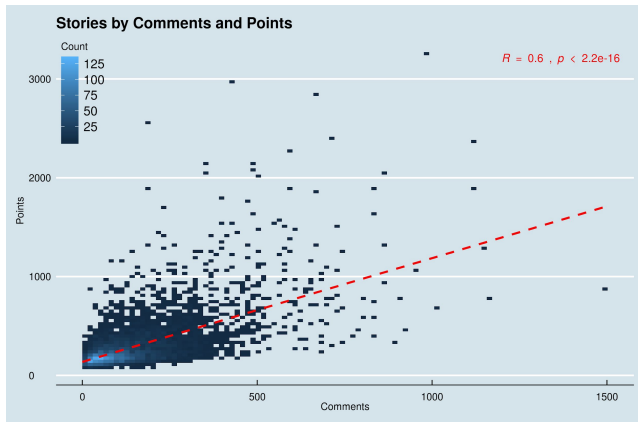


An abstract graphic featuring a large, irregular white shape in the center, surrounded by a thick, dark blue ring. The background is a light gray.

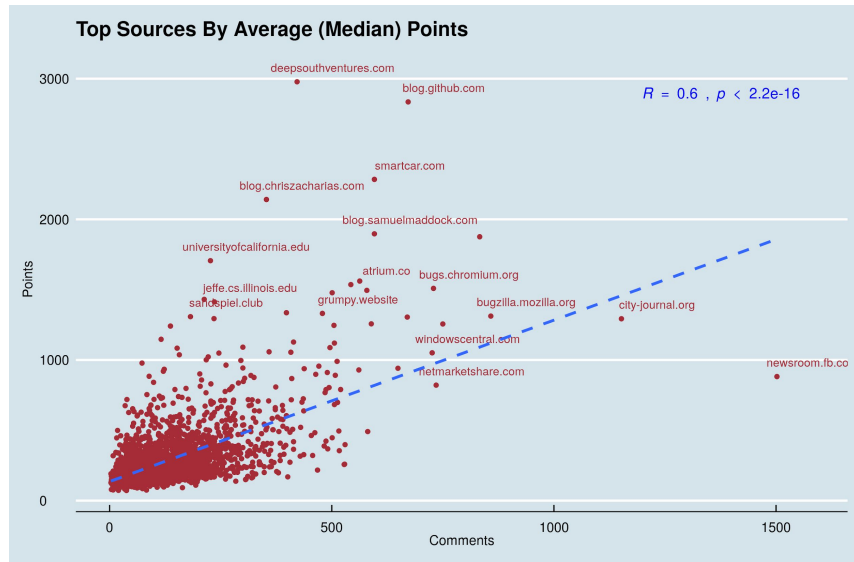
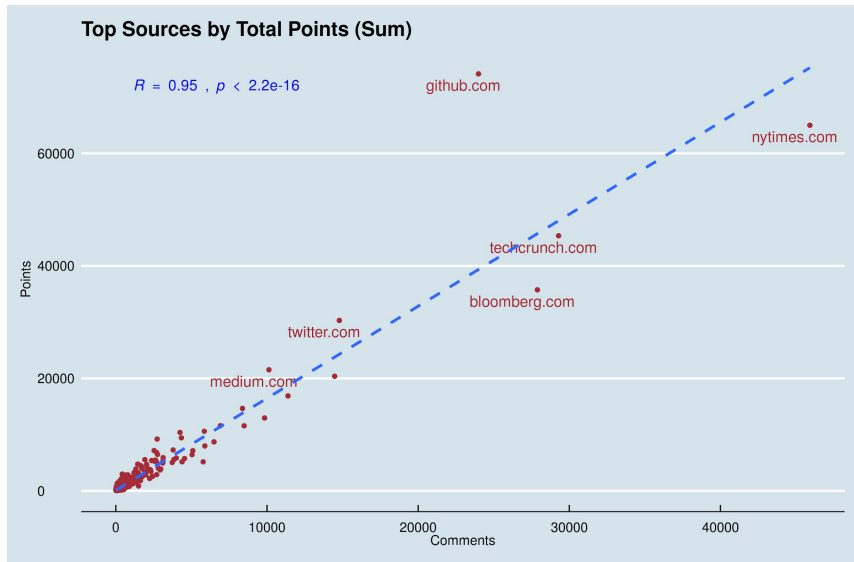
## **Descriptive Analysis**

## Distribution by Points and Comments

- Perhaps unsurprisingly, points and comments are **moderately to strongly correlated** with a Pearson's R of 0.6
- This correlation decreases slowly, when grouped by rank
- High density at the zero-axis
- Gentle horizontal and vertical skews - sometimes users comment without upvoting and vice-versa

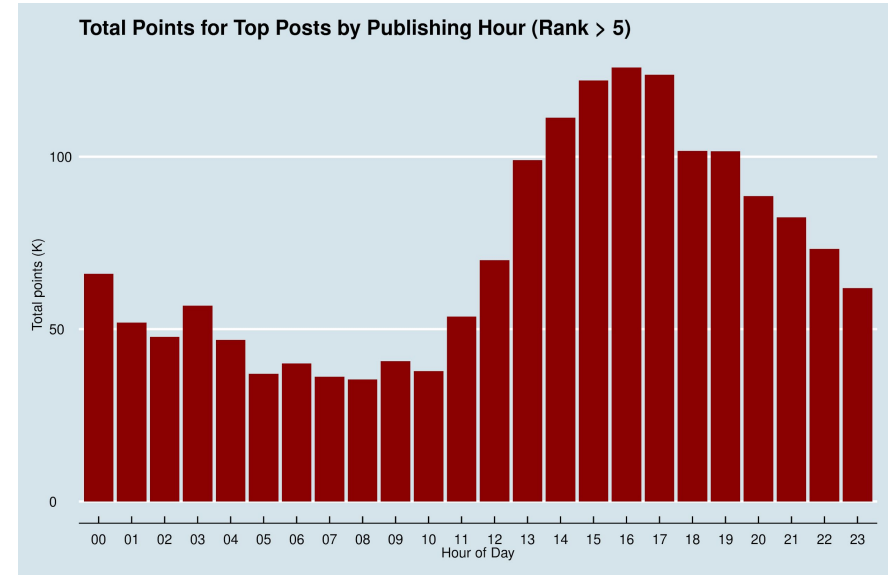
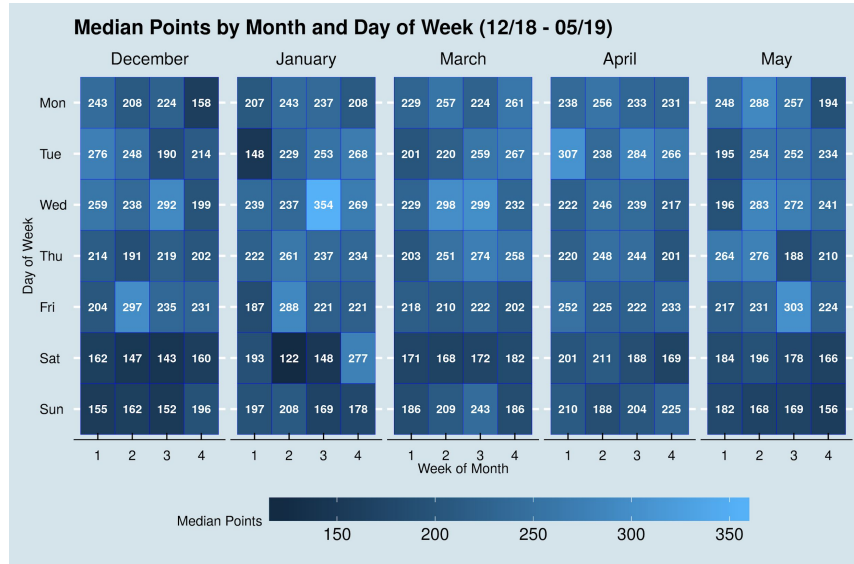


## Sources by Points and Comments



Source by sum gives a clear picture of the top sources -  
but by median, we see some sources are *one-hit-wonders*

## Date & Time of Day



Stories Published on Weekdays and between 2pm and 7pm tend to be assigned more points





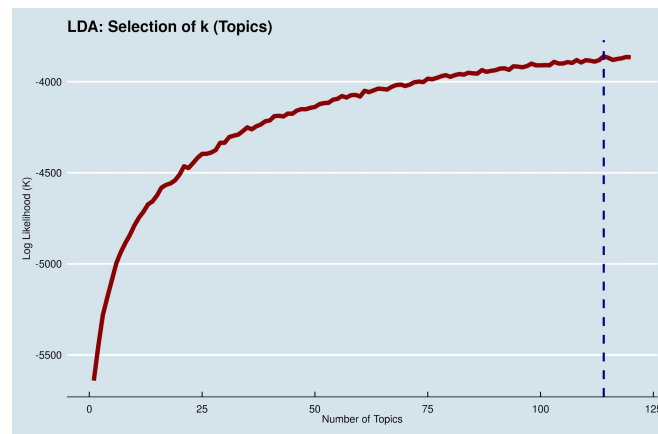
## **Topic Analysis**

## Detecting Topics

- We perform punctuation removal and stemming on our corpus
- *Gibbs Sampling* is used to find the conditional probability distribution of a word's topic assignment
- Our model is run at 700 iterations, discarding the first 200 iterations to increase accuracy
- We re-train the model for  $1 < K < 121$ , calculating the Maximum Log Likelihood at each iteration - selecting k when the value MLL begins to taper off
- Ideal K for our model is 114 topics (detected)

Document Term Matrix	
Documents	5,939
Terms	7,947
Sparsity	99%

TABLE I  
DOCUMENT TERM MATRIX FOR LDA TRAINING



## Topics Detected

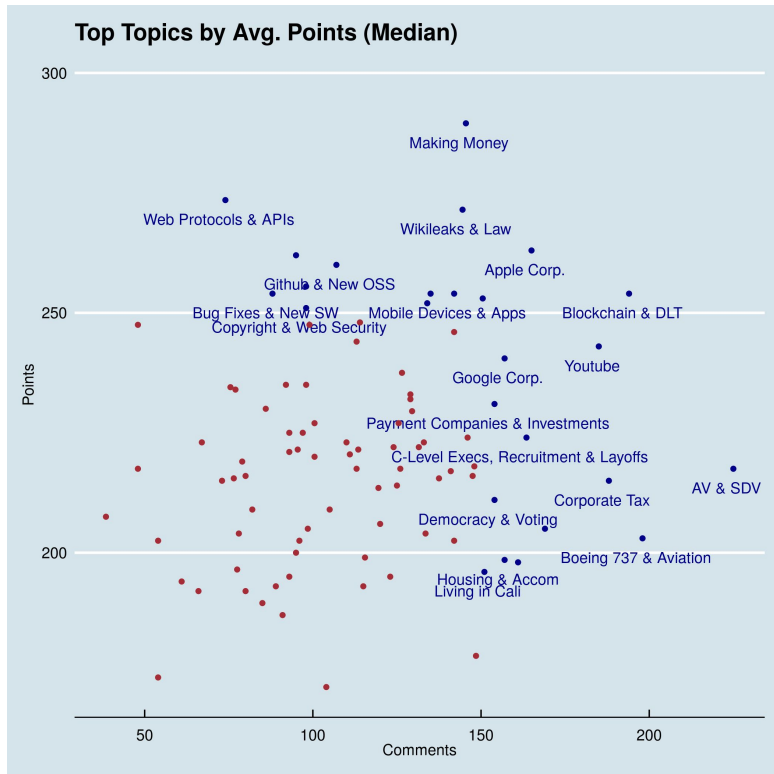
- A total of 114 topics were detected
- We evaluated each topic, assigning an annotation to each one
- The topics are of generally high quality, 15% of topics were identified to be noisy.

Topic Id	Keywords	Human-Readable Annotation
1	point view channel youtub content	Youtube
2	flight plane air boe pilot	Boeing 737 & Aviation
3	countri india germani europ world	Geopolitics
4	phone devic call android smartphon	Mobile Devices & Apps
5	life friend mind world advic	Self Improvement
6	rust webassembl red swift runtim	Rust Lang.
7	materi glass plastic wast wood	Environmental Concerns
8	earth moon land planet star	Space Exploration
9	compani fund investor capit founder	VC, IPOs & Money
10	editor edit guid emac studio	IDEs
11	money peopl dollar million thousand	Making Money
12	peopl decis lot problem fact	Quitting Jobs & Toxic SV Culture
13	ve lot ll note	Noisy Topic 1
14	perform intel cpu processor core	CPUs, Performance & Advances
15	market busi industri compani sale	Corporate Revenue & Transactions

TABLE II  
EXAMPLES OF HUMAN-READABLE ANNOTATIONS ASSIGNED TO  
LDA TOPICS

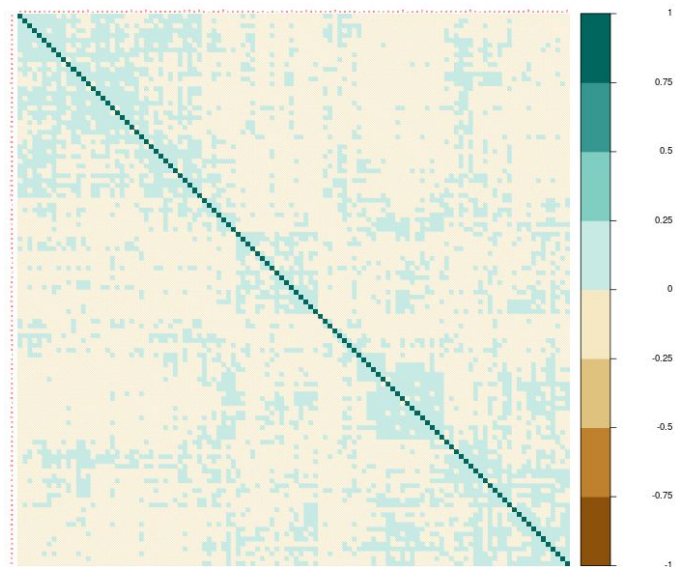
## Top Topics

Annotation	Story Headline
Hiring & Careers	Absolute truths I unlearned as junior developer (monicalent.com)
Hiring & Careers	I interviewed at six top companies in Silicon Valley in six days (blog.usejournal.com)
Hiring & Careers	When hiring senior engineers, youre not buying, youre selling (hiringengineersbook.com)
Making Money	How to Be Successful (blog.samaltman.com)
Making Money	Firefox desktop market share now below 9% (netmarketshare.com)
Making Money	Open Source Doesnt Make Money Because It Isnt Designed to Make Money (www.ianbicking.org)
Web Protocols & APIs	Remote Code Execution on Most Dell Computers (d4stiny.github.io)
Web Protocols & APIs	HTTP headers for the responsible developer (www.twilio.com)
Web Protocols & APIs	HTTP/3 explained (http3-explained.haxx.se)
Wikileaks & Law	Julian Assange arrested in London (www.bbc.co.uk)
Wikileaks & Law	U.S. Supreme Court Puts Limits on Police Power to Seize Private Property (www.nytimes.com)
Wikileaks & Law	If Software Is Funded from a Public Source, Its Code Should Be Open Source (www.linuxjournal.com)
Apple Corp.	Spotify to Apple: Time to Play Fair (www.timetoplayfair.com)
Apple Corp.	FaceTime bug lets you hear audio of person you are calling before they pick up (9to5mac.com)
Apple Corp.	Apple Sign In (techcrunch.com)

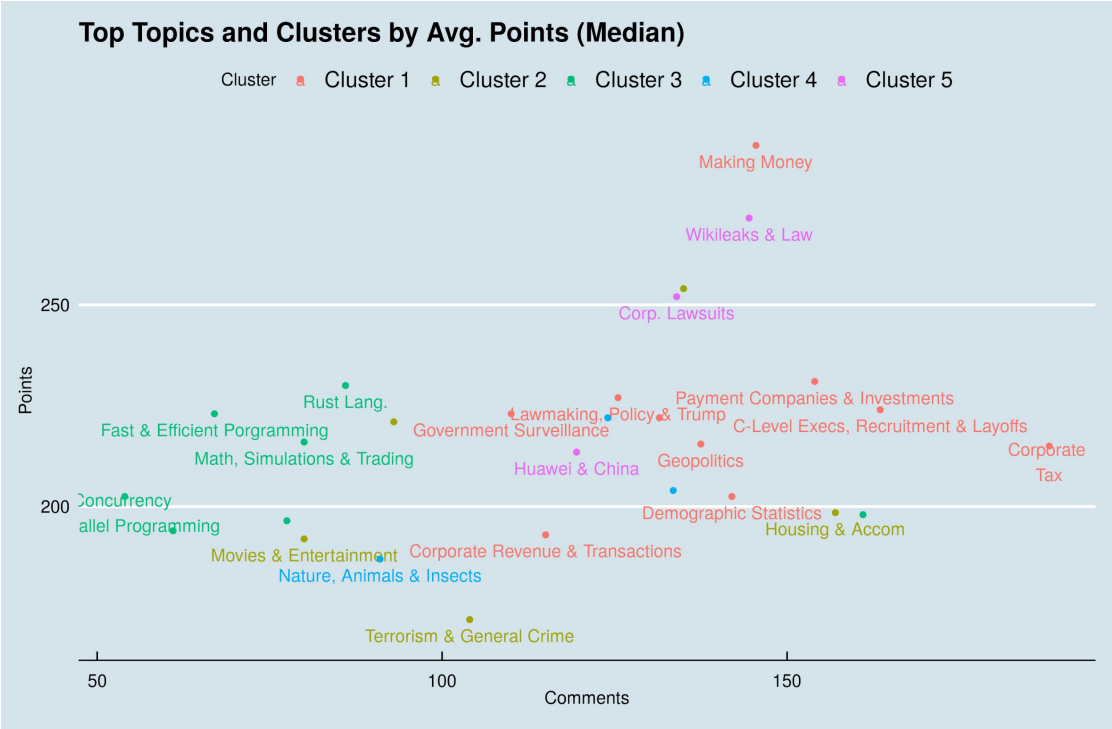


## Covariance Matrix: Do certain topics occur more often than others?

- We observe that the concept of a topic does not occur in isolation - topics are frequently inter-related.
- The most frequently occurring topics are those which occur most frequently together.
- We build a document-topic probability covariance matrix, displaying the covariance of two topics occurring together.
- The matrix is clustered hierarchically - we can see that topic clusters clearly exist



# Topic Clusters: Do certain topics occur more often than others?



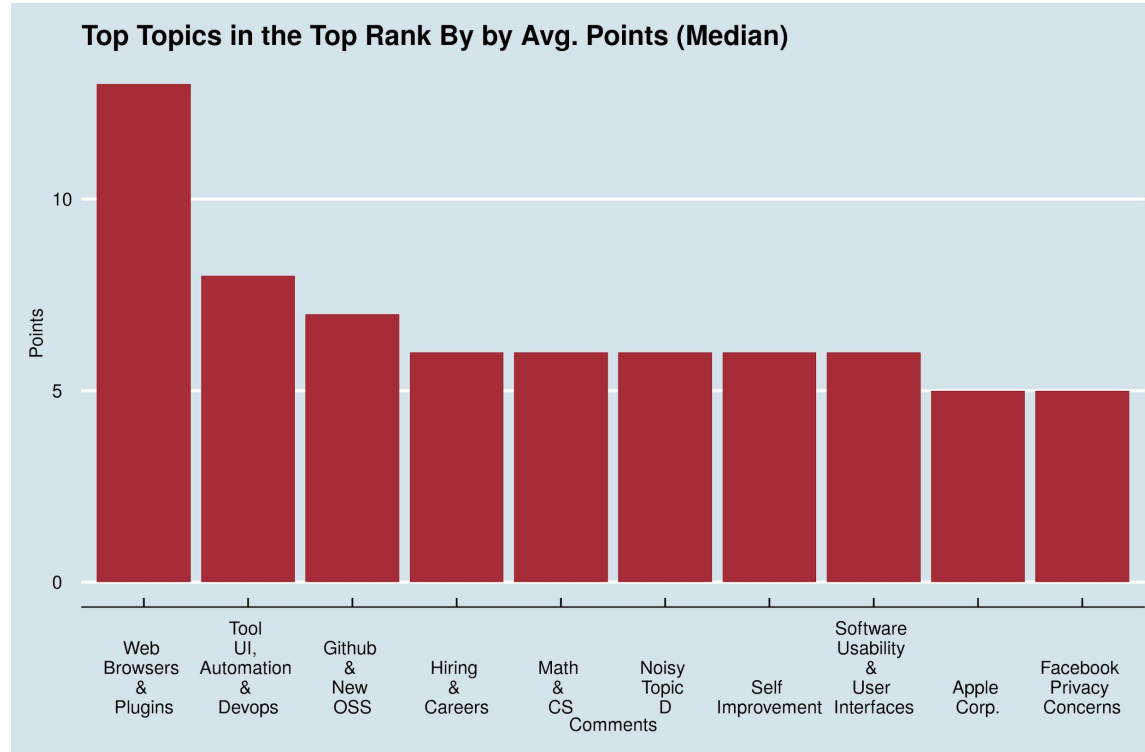
Topic Id	Topic Annotation	Cluster Name
11	Making Money	Cluster 1
15	Corporate Revenue & Transactions	Cluster 1
37	Corporate Tax	Cluster 1
23	C-Level Execs, Recruitment & Layoffs	Cluster 1
3	Geopolitics	Cluster 1
17	Lawmaking, Policy & Trump	Cluster 1
43	Payment Companies & Investments	Cluster 1
9	VC, IPOs & Money	Cluster 1
50	Government Surveillance	Cluster 1
30	Demographic Statistics	Cluster 1
51	Noisy Topic G	Cluster 1
48	Housing & Accom	Cluster 2
22	Japanese Tech.	Cluster 2
35	Problems & The Past	Cluster 2
21	Noisy Topic B	Cluster 2
20	Movies & Entertainment	Cluster 2
12	Quitting Jobs & Toxic SV Culture	Cluster 2
27	Terrorism & General Crime	Cluster 2
47	Math, Simulations & Trading	Cluster 3
53	Functional Languages	Cluster 3
44	Noisy Topic E	Cluster 3
6	Rust Lang.	Cluster 3
31	Noisy Topic C	Cluster 3
40	Ruby Lang.	Cluster 3
33	Interpreted Prog. Languages	Cluster 3
57	Concurrency & Parallel Programming	Cluster 3
26	Fast & Efficient Programming	Cluster 3
77	Nature, Animals & Insects	Cluster 4
113	Family Issues & Happiness	Cluster 4
94	Chess & Games & Board Games	Cluster 4
85	Noisy Topic K	Cluster 4
110	Noisy Topic Q	Cluster 5
89	Huawei & China	Cluster 5
58	Wikileaks & Law	Cluster 5
71	Corp. Lawsuits	Cluster 5

TABLE IV  
CLUSTERS EXTRACTED FROM COVARIANCE MATRIX

## Rank Correlation: Are specific topics over-represented in the top post?

- We pose the null hypothesis  $H_0$ :
  - *Achieving the top rank from within the front page is unrelated to the story topic*
- We perform a **Chi-Squared test** between two categorical values: topic name and binned rank (Top Rank/Not Top Rank):
  - **The top rank is highly dependent on the topic p-value of 0.002634**
- We perform a One-Way ANOVA test between the numeric rank (*ordinal*) and the topic name (*categorical*):
  - **The top rank is highly dependent on the topic with  $\text{Pr}( > F )$  2.97-e09**
- We **reject** the null hypothesis, **the top rank is highly dependent on the topic.**

## Top Topics within the Top Post







## **Conclusion**

## Key Findings

- Points and comments are moderately to strongly correlated with a Pearson's R of 0.6
- Stories posted between 2pm and 7pm tend to rank higher than others.
- Certain topics tend to be over-represented on the front page such as:
  - These include topics related to companies and revenue growth (such as making money, recruitment)
  - Programming Niches (Ruby/Rust Programming, Concurrency and Parallel Programming)
- There is strong evidence which states that achieving the top rank is dependent on the topic ( $p=0.002634$ )
  - Stories on “Web Browsers & Plugins” are over-represented within the top rank.

## Evaluation

- The detected topics are of a high quality mainly due to our decision represent documents as collections of *Noun Phrases and Verb Phrases*
- We observe a very low p-value in our top rank analysis, noting that the chi-squared test may be unreliable due to low frequency counts. A Fisher's exact test rendered similar results
  - For this reason we also performed a One-Way ANOVA on the full rank value
- It may be interesting to further investigate the temporal aspect of the data