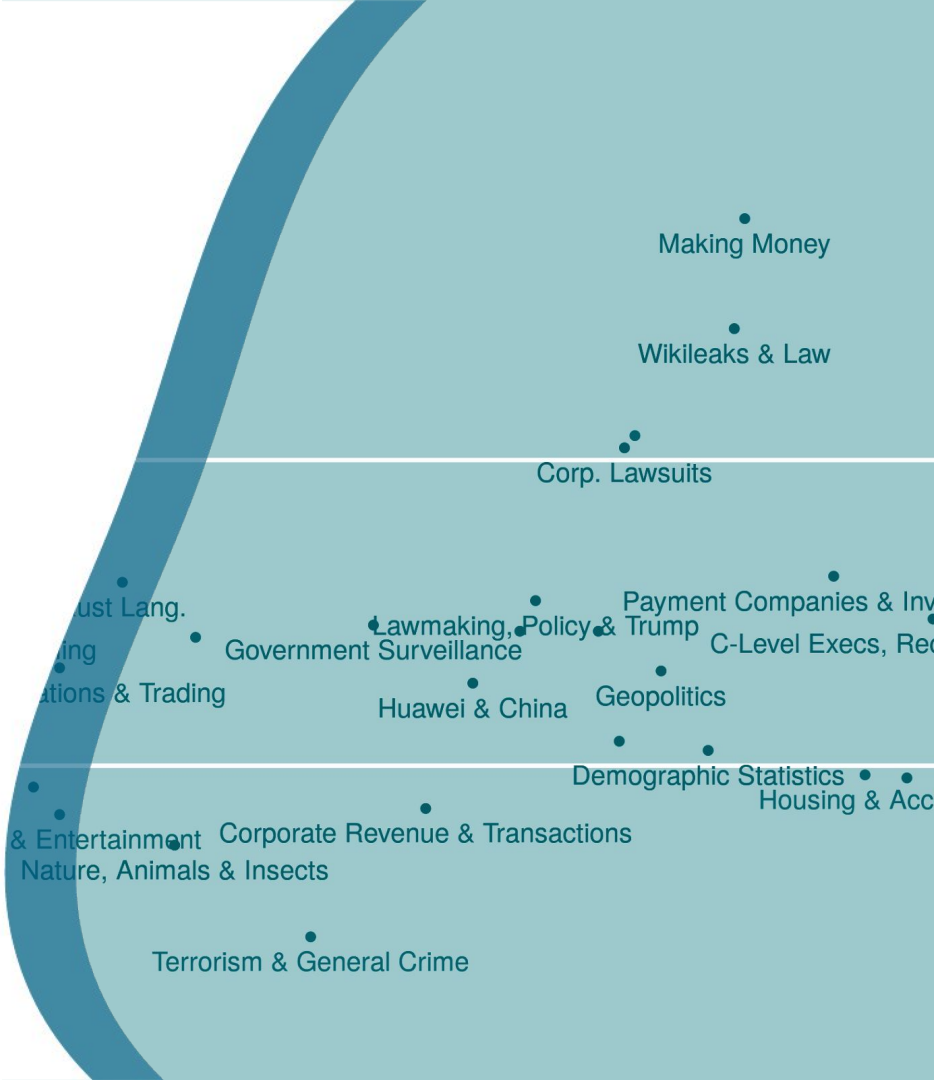
L-Università ta' Malta

Analyzing Topic Prevalence in Popular Hacker News Stories

ICS5115 - Statistics for Data Scientists

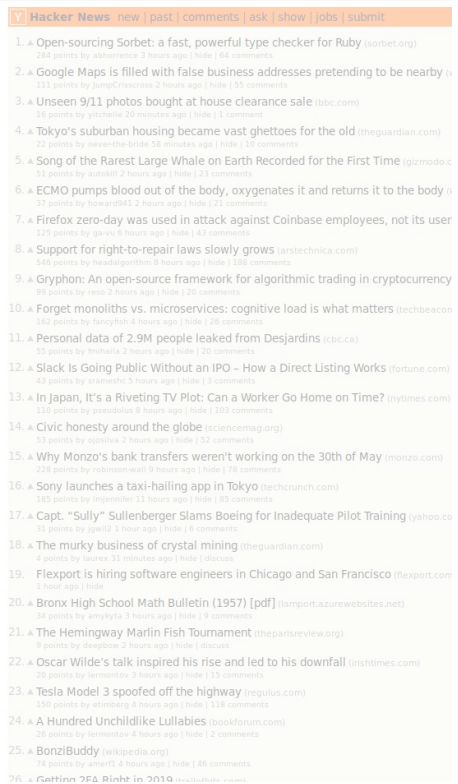
Jake J. Dalli
jake.dalli.10@um.edu.mt
MSc. Artificial intelligence
Faculty of ICT



Jake J. Dalli (jake.dalli.10@um.edu.mt)

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*



Jake J. Dalli (jake.dalli.10@um.edu.mt)

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
https://hckrnews.com/
Several News Websites

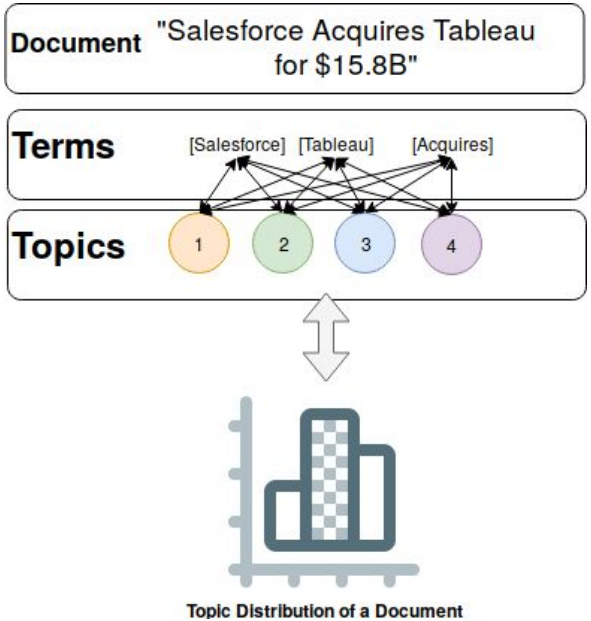
Source Datasets
hackernews_board.csv
hackernews_board_content.csv

Jake J. Dalli (jake.dalli.10@um.edu.mt)

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*

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

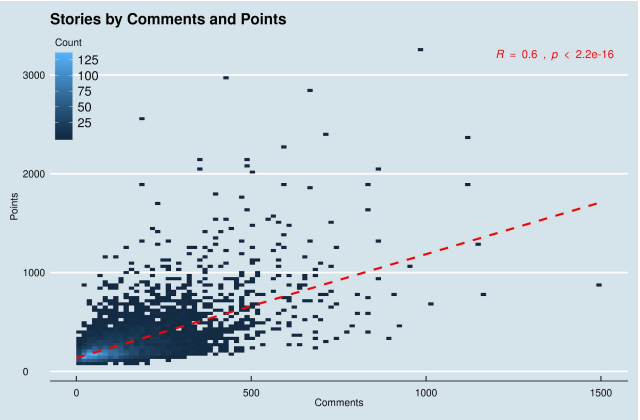


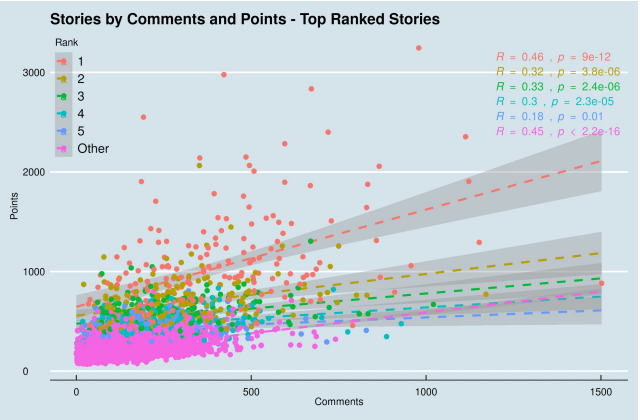
Descriptive Analysis

Jake J. Dalli (jake.dalli.10@um.edu.mt)

Distribution by Points and Comments

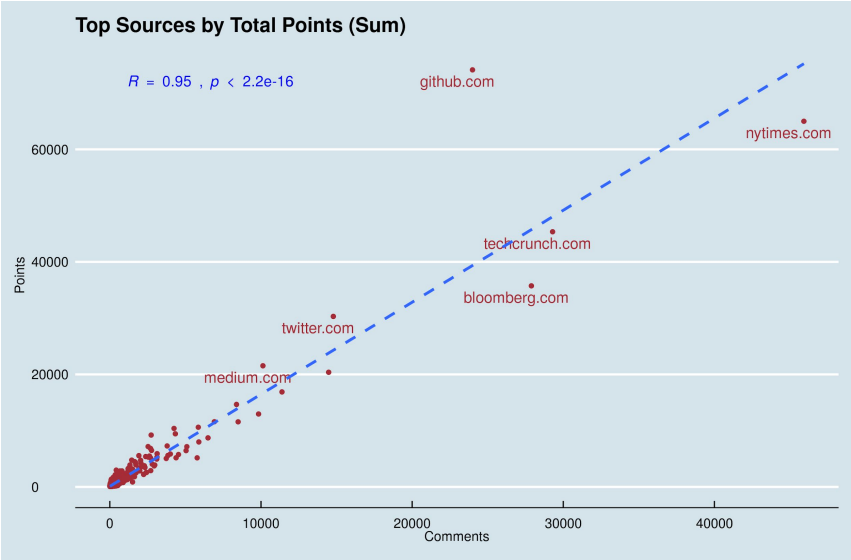
- Perhaps unsurprisingly, points and comments are **moderately to strongly correlated** with a Pearson's R of 0.6
- This correlation decreases slowly, when grouped by rank
- High density at the zero-axis
- Gentle horizontal and vertical skews - sometimes users comment without upvoting and vice-versa

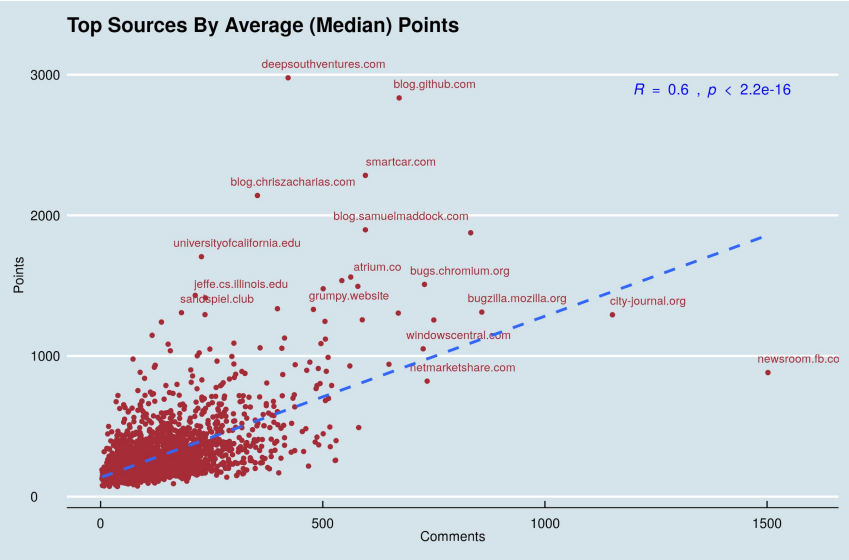




Jake J. Dalli (jake.dalli.10@um.edu.mt)

Sources by Points and Comments

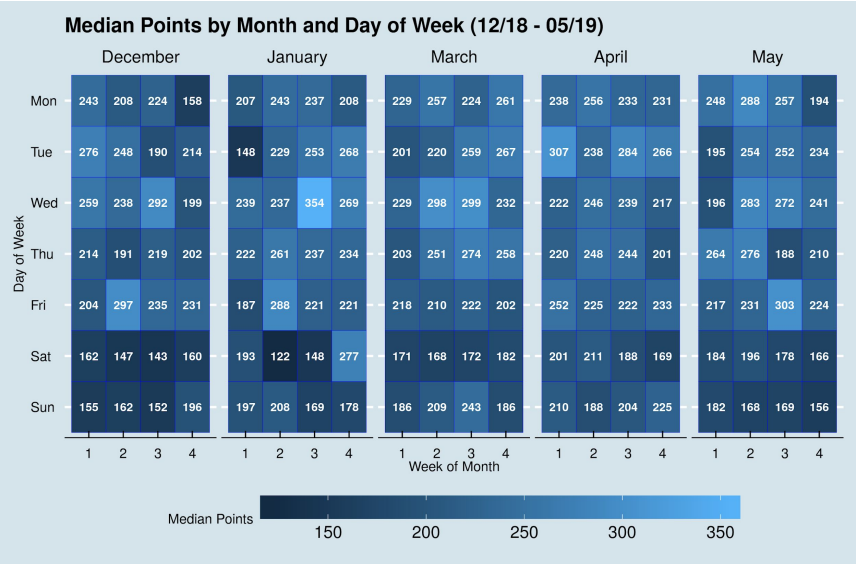


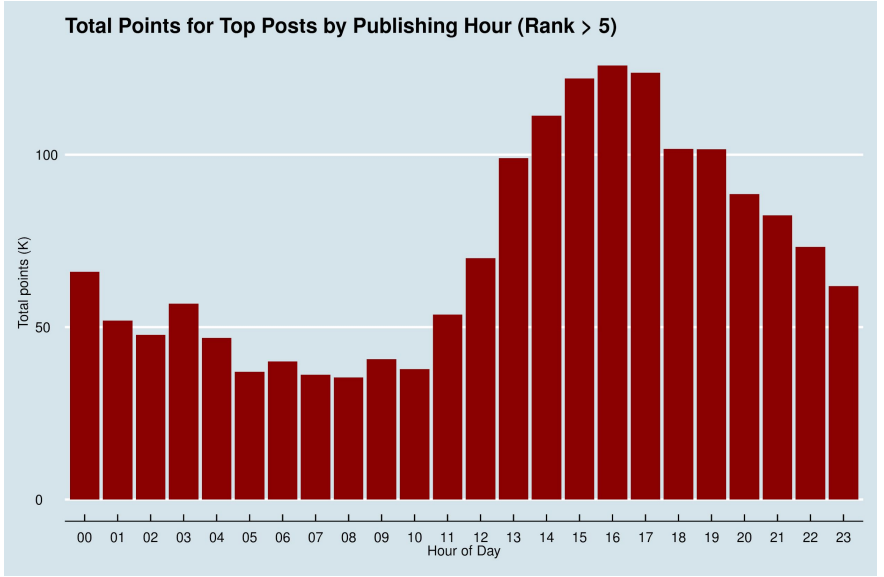


Source by sum gives a clear picture of the top sources -
but by median, we see some sources are *one-hit-wonders*

Jake J. Dalli (jake.dalli.10@um.edu.mt)

Date & Time of Day





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

Topic Analysis

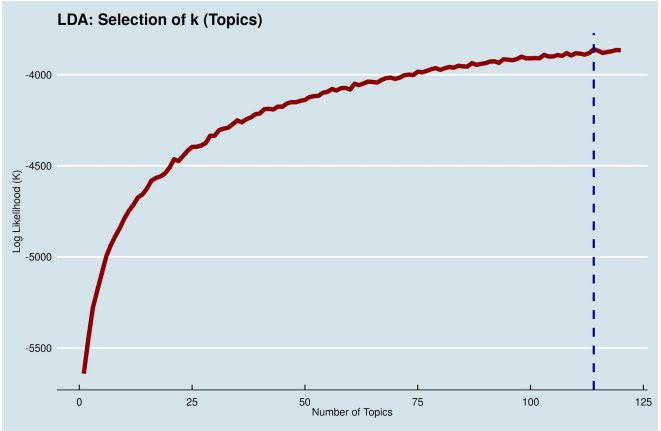
Jake J. Dalli (jake.dalli.10@um.edu.mt)

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



Jake J. Dalli (jake.dalli.10@um.edu.mt)

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.

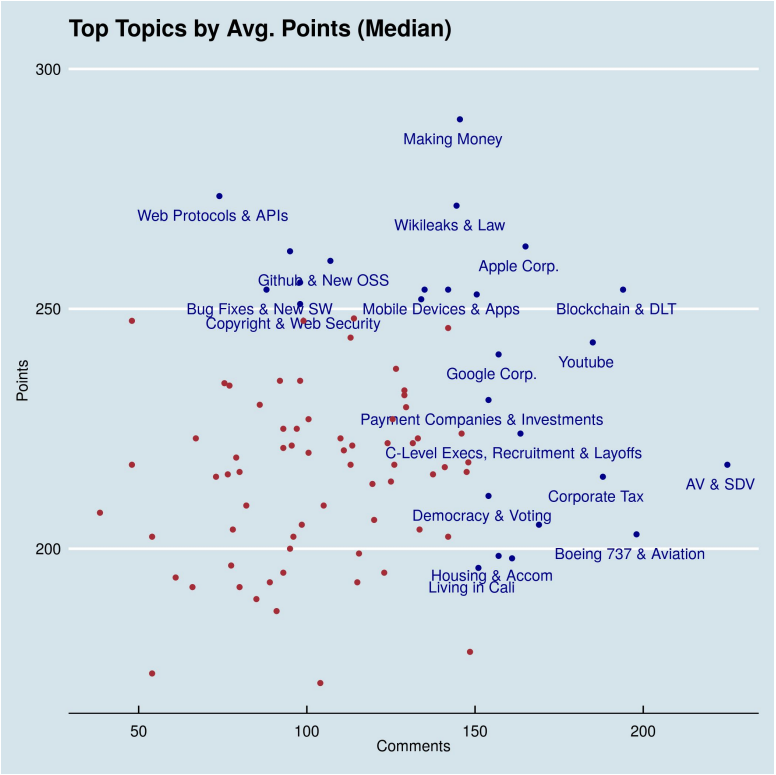
Topic Id	Keywords	Human-Readable Annotation
1	point view channel youtub content	Youtube
2	flight plane air boe pilot	Boeing 737 & Aviation
3	countri india germani europ world	Geopolitics
4	phone devic call android smartphon	Mobile Devices & Apps
5	life friend mind world advic	Self Improvement
6	rust 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

Jake J. Dalli (jake.dalli.10@um.edu.mt)

Top Topics

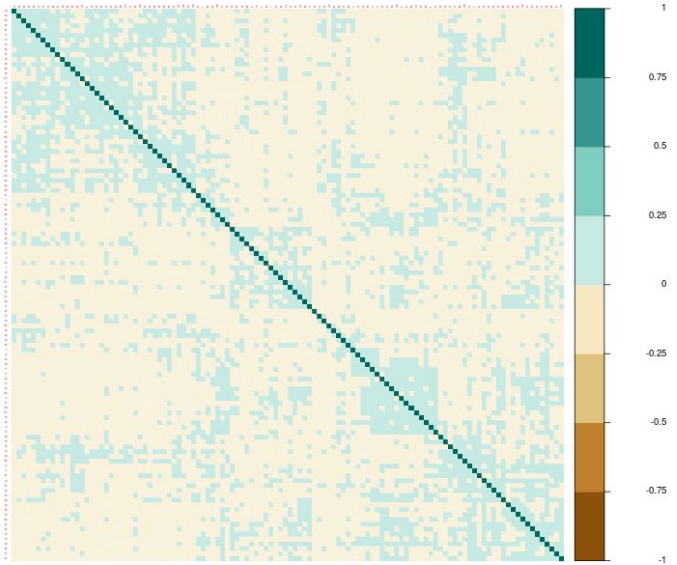
Annotation	Story Headline
Hiring & Careers	Absolute truths I unlearned as junior developer (monicalent.com)
Hiring & Careers	I interviewed at six top companies in Silicon Valley in six days (blog.usejournal.com)
Hiring & Careers	When hiring senior engineers, youre not buying, youre selling (hiringengineersbook.com)
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)



Jake J. Dalli (jake.dalli.10@um.edu.mt)

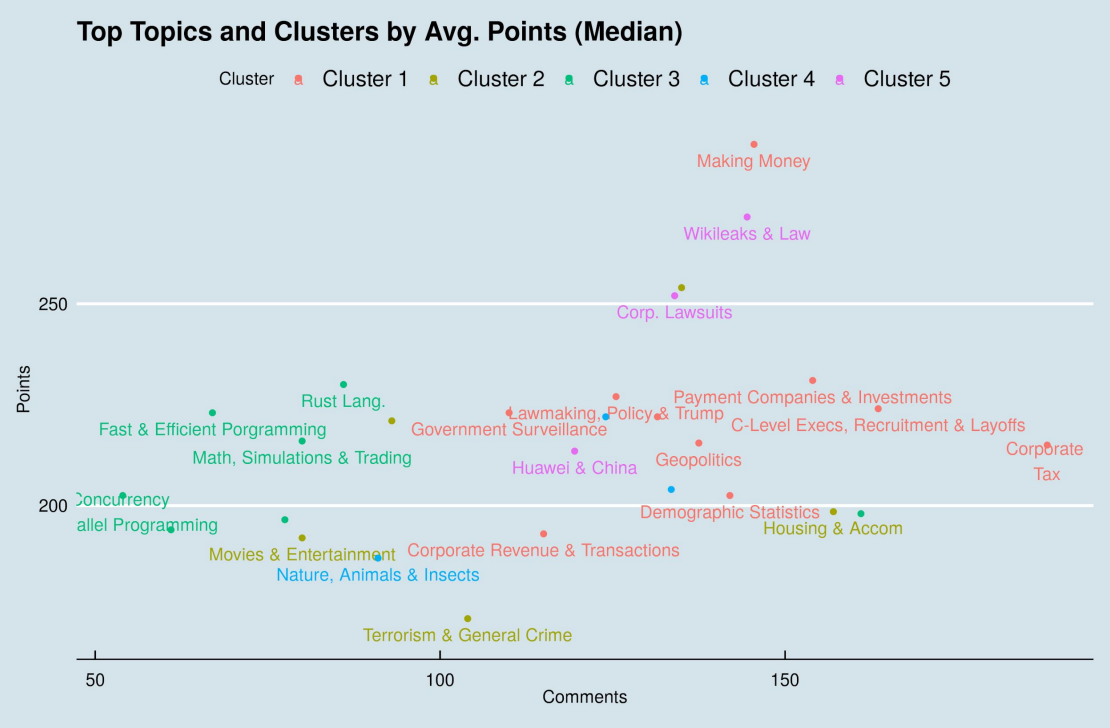
Covariance Matrix:
Do certain topics occur more often than others?

- We observe that the concept of a topic does not occur in isolation - topics are frequently inter-related.
- The most frequently occuring 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



Jake J. Dalli (jake.dalli.10@um.edu.mt)

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	Nosiy 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 Lagnuages	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 Porgramming	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

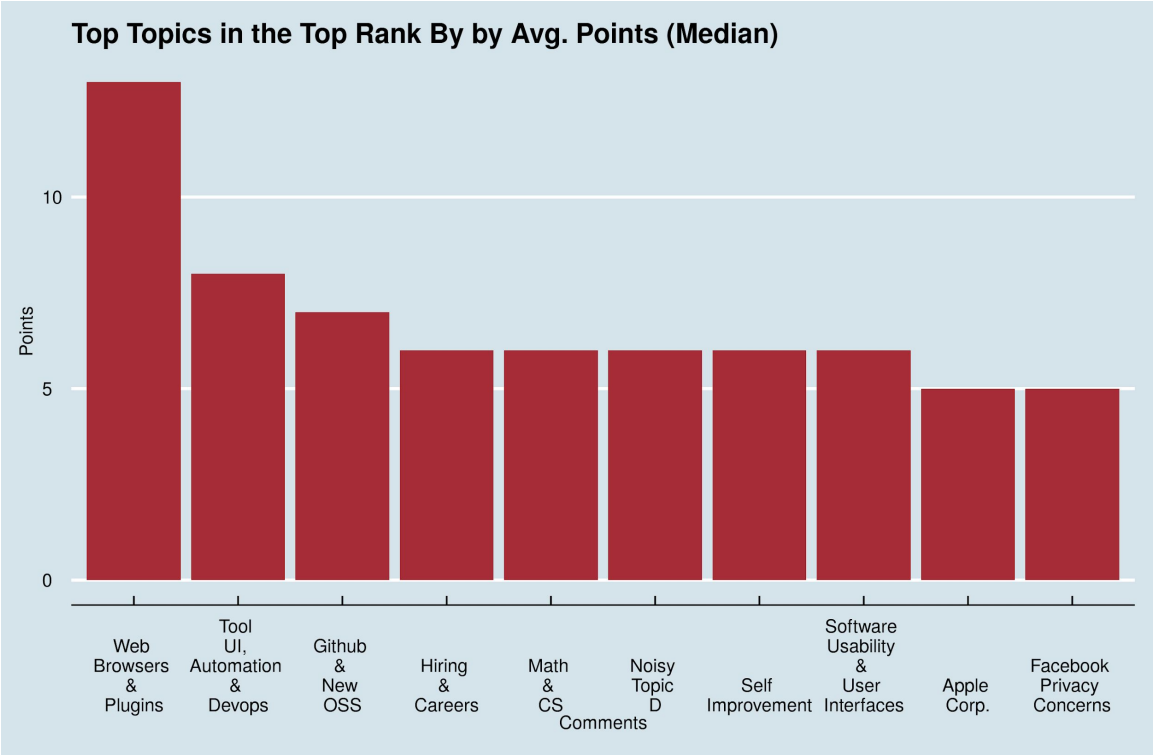
Jake J. Dalli (jake.dalli.10@um.edu.mt)

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 Pr(>F) 2.97-e09**
- We **reject** the null hypothesis, **the top rank is highly dependent on the topic.**

Jake J. Dalli (jake.dalli.10@um.edu.mt)

Top Topics within the Top Post



Conclusion

Jake J. Dalli (jake.dalli.10@um.edu.mt)

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.

Jake J. Dalli (jake.dalli.10@um.edu.mt)

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