Knowledge Assistant for Comment Systems

Capstone Presentation – A Case Study on HODINKEE.com

May 7, 2020

Hamza Masood

LinkedIn | Data Science – Flatiron School | GitHub

Contents

- Why is this interesting?
- Process Steps
- Ideas for the future

Why is this interesting?

To Hodinkee

- Better user engagement → more time spent on hodinkee.com → Higher conversion rate for the Hodinkee Shop
- Build a better community than competitors

To a General Audience

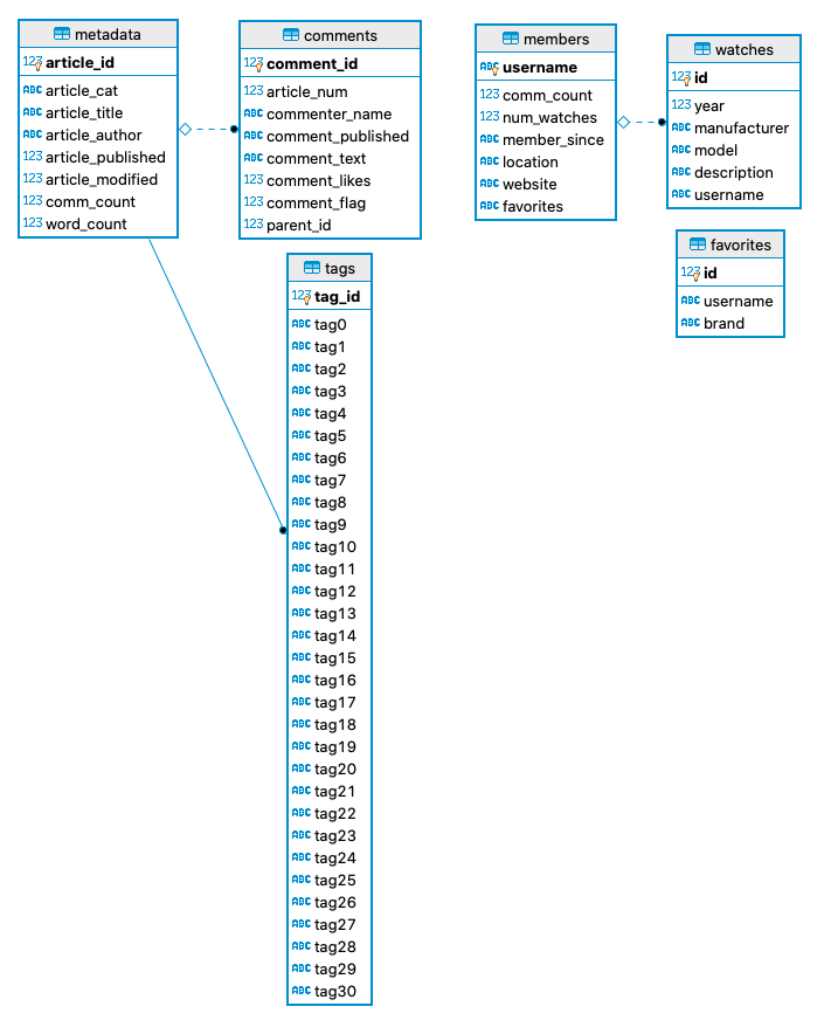
- Lots of similar online communities with highly engaged users: other online watch clubs, tech review sites, wine clubs, gaming reviews, and so on

To Me

- I love watches
- I got to build a dataset from scratch

Code is privately stored on GitHub since this is a proprietary dataset

Methodology Details



SQLite database schema diagram

Website snapshot taken on Apr 5

Total scraped data >1.75GB

Python libraries used:

- pandas
- numpy
- beautifulsoup4
- sqlite
- matplotlib
- statsmodels
- datetime
- colour

metadata	*6,501 rows*
tags	*6,501 rows*
comments	*137,495 rows*
members	*13,355 rows*
favorites	*17090 rows*
watches	*13490 rows*

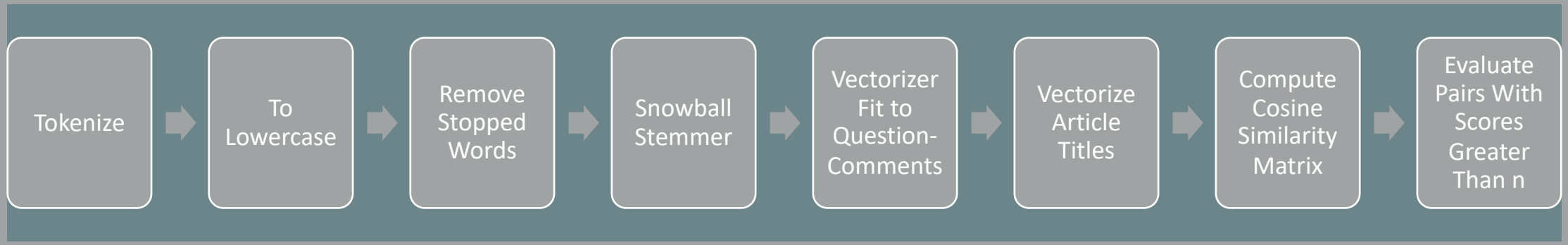
6,501 article html pages
6,501 article text files
13355 user html pages

Some high level numbers

	Articles	Comments
Disqus Comments System *(2010 – 2016)*	3,636	
New Comments System *(2017 – present)*	2,865	137,495

Article Categories (not specified counted as one category)	**66**
Median article word count	**457**
90th percentile article word count	**1,381**
Comments posted in reply to an existing comment	**36,183**
Comments by Hodinkee staff	**3,129**
Duplicate comments posted by user in error	**315**
Most commented article: "Friday Live Episode 16: What Three Watches Would You Buy With $15,000?"	**717**
Median number of comments per article	**32**
Articles with zero comments under the new system	**10**
Comments with a "?" in the comment text	**25,640**

FSM



Issues Found

- Recommended article can't be the same as the article on which the comment was written.
- Recommended article can't have a publish date greater than the comment publish date.
- Types of questions:
  - Specific (who is, what is, why is, how is)
  - Not answerable by an article:
    - Replying to another comment with a question
    - contextually specific to the article itself
    - Sarcasm
    - "?!?!"
    - rhetorical question answered in the same comment

Increasing Relevance

Evaluation Scale

1. Perfect
2. Thematically Relevant
3. Related But Different
4. Irrelevant

Results

Only 49 comments with CS scores > 0.8, most were evaluated as NOT irrelevant

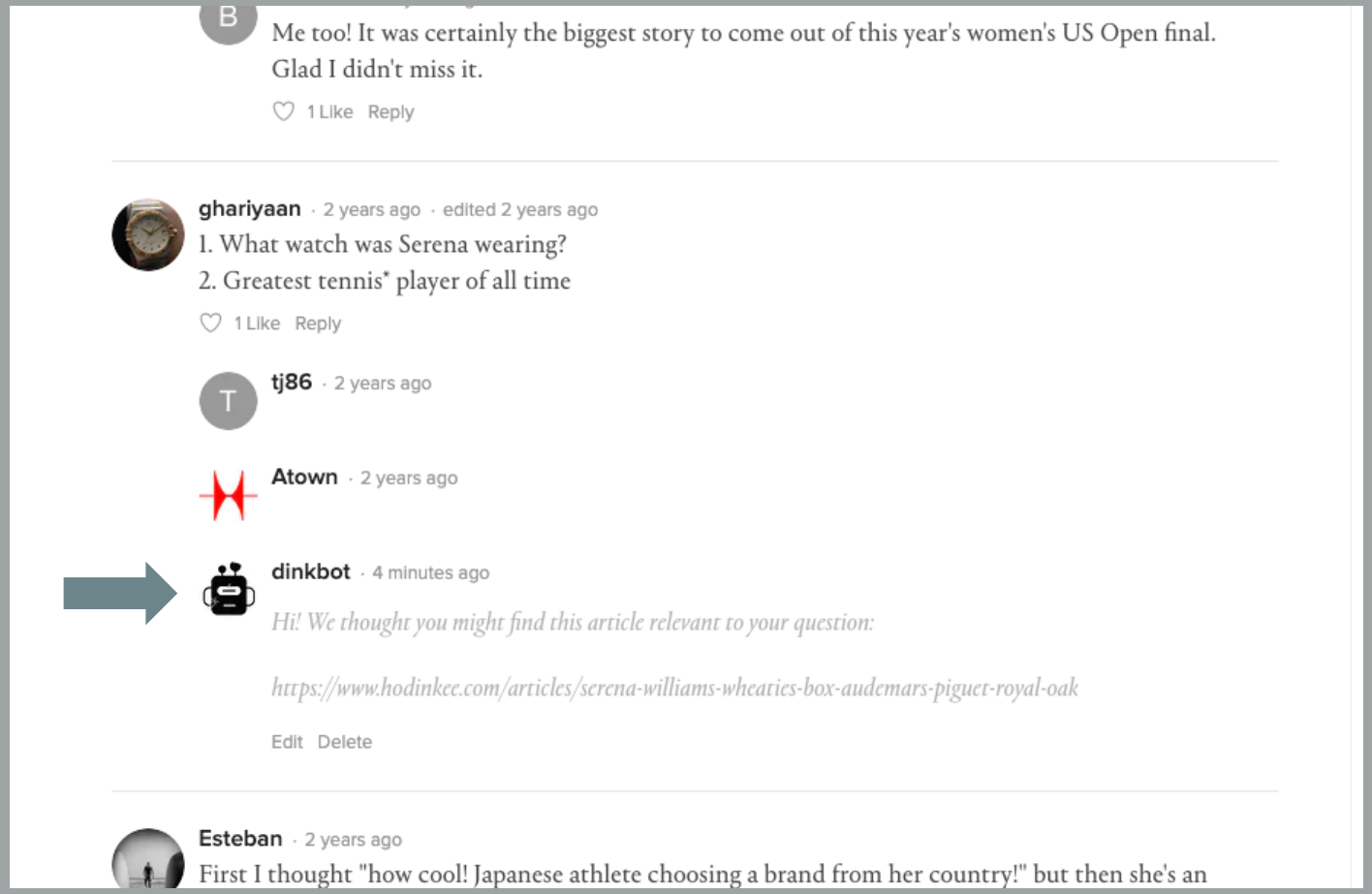
Digging into the questions

	Articles	Comments
Disqus Comments System *(2010 – 2016)*	3,636	
New Comments System *(2017 – present)*	2,865	137,495

Article Categories (not specified counted as one category)	**66**
Median article word count	**457**
90th percentile article word count	**1,381**
Comments posted in reply to an existing comment	**36,183**
Comments by Hodinkee staff	**3,129**
Duplicate comments posted by user in error	**315**
Most commented article: "Friday Live Episode 16: What Three Watches Would You Buy With $15,000?"	**717**
Median number of comments per article	**32**
Articles with zero comments under the new system	**10**
Comments with a "?" in the comment text	**25,640**

Round 2

Ideas for the future



- Live site deployment
- Article corpus expanded from Hodinkee to the Watchville publisher family and beyond

Questions?

Sources & Kudos

- Hodinkee.com

- Erin Hoffman
- Greg Damico
- Brian Mcgarry

Appendix

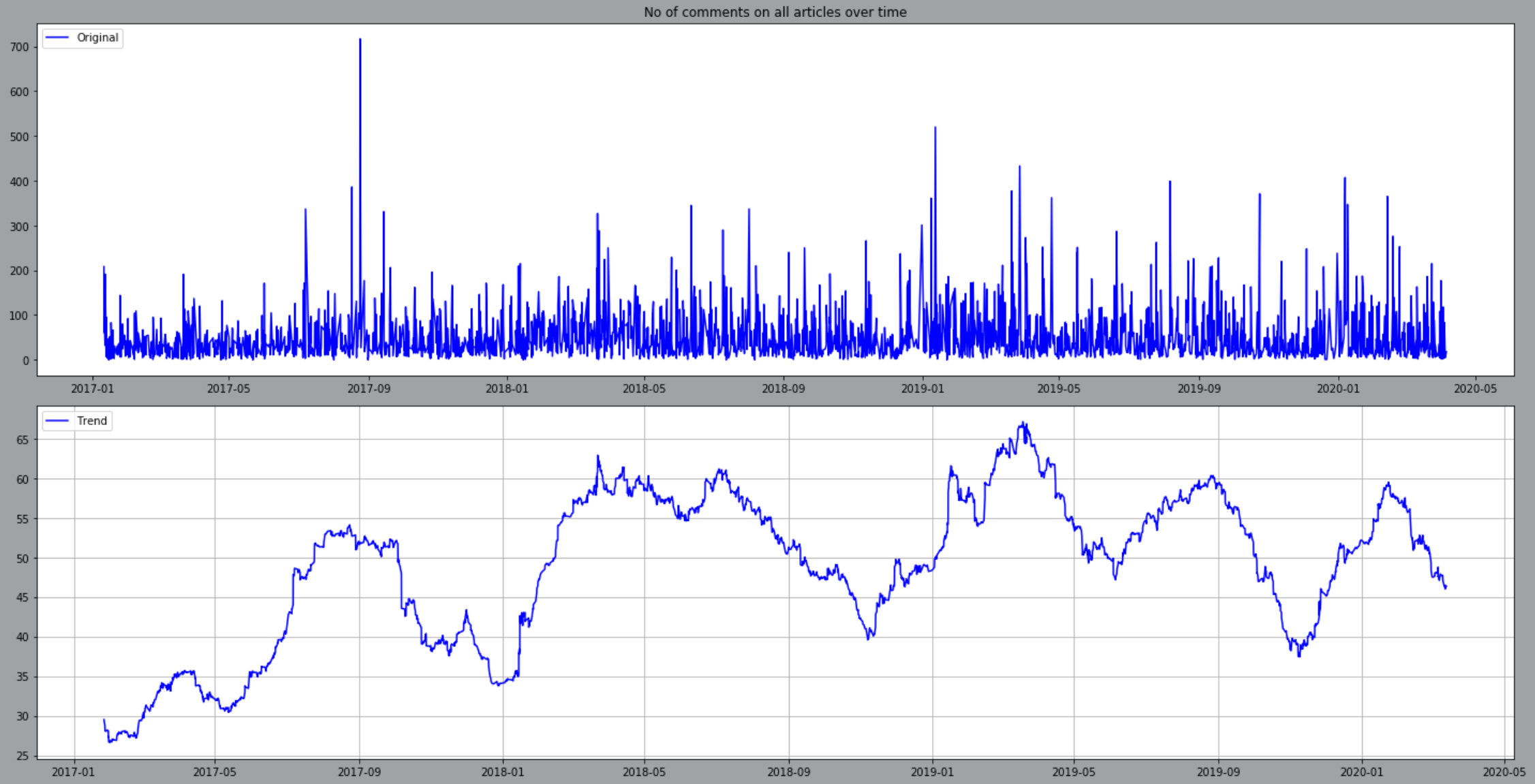
All-up Assumptions

- Subjectivity of evaluation scale

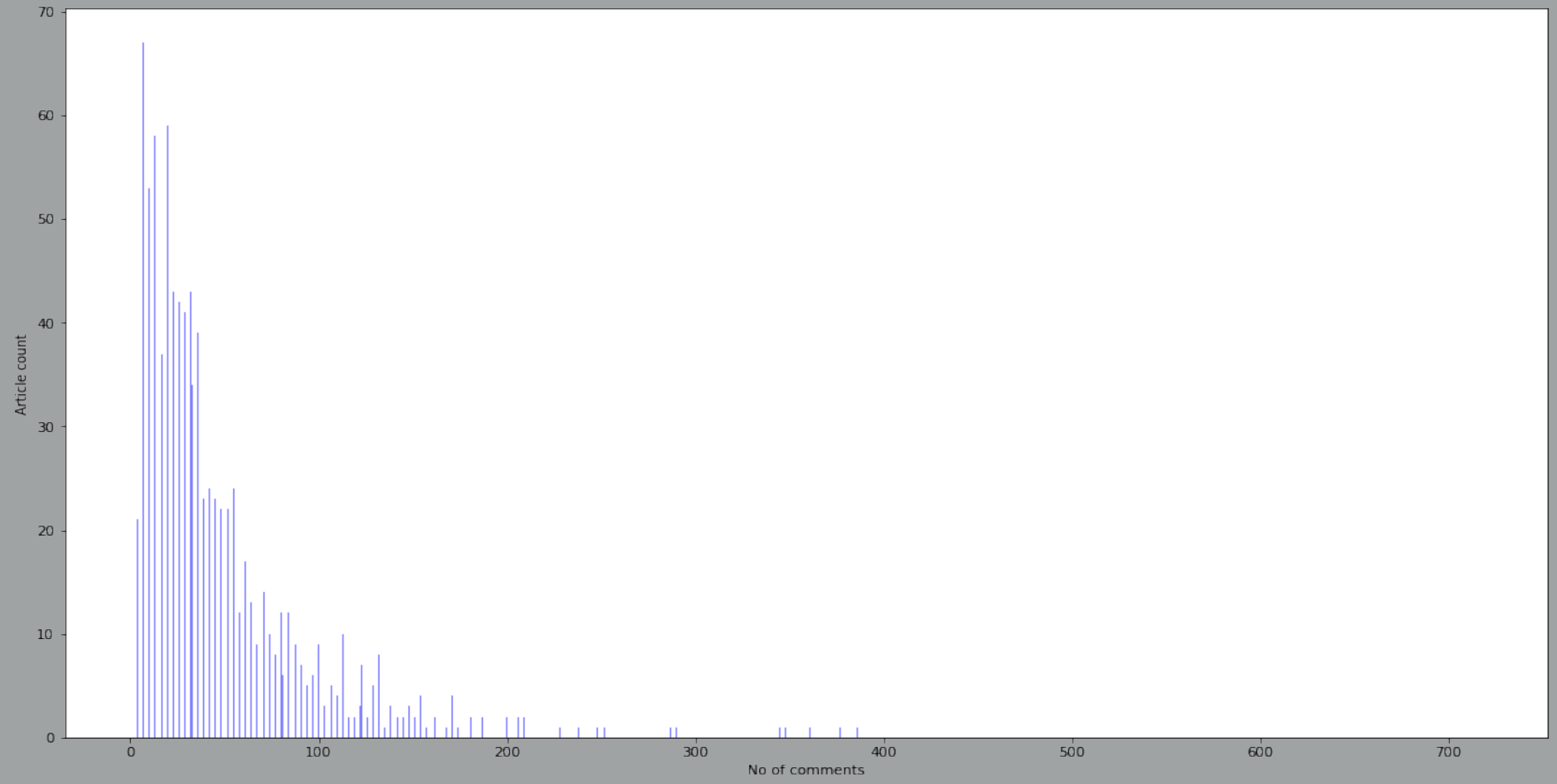
- Website snapshot was taken at end of day, Saturday April 4, 2020; all new content since then not included in the analysis
- User profile snapshots taken on Friday April 17, 2020

- Disqus comments not analyzed
- Articles only, no special coverage included; articles with more than one author in the byline default to taking the first one as the only author
- For comments in reply to a deleted comment, the parent comment is assumed to be the next comment up in the tree, or no parent if the deleted comment was itself a parent comment.

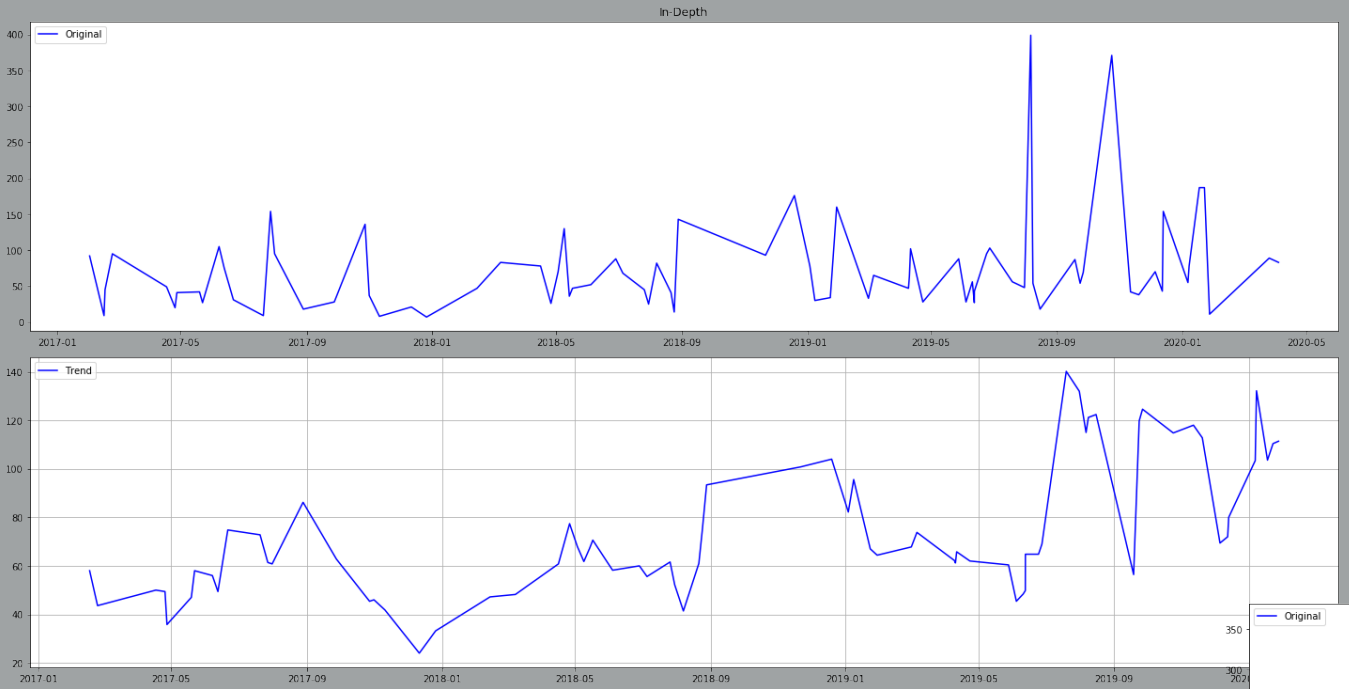
All comments over time: seasonal, cyclical trend



Distrubution of comment count over articles

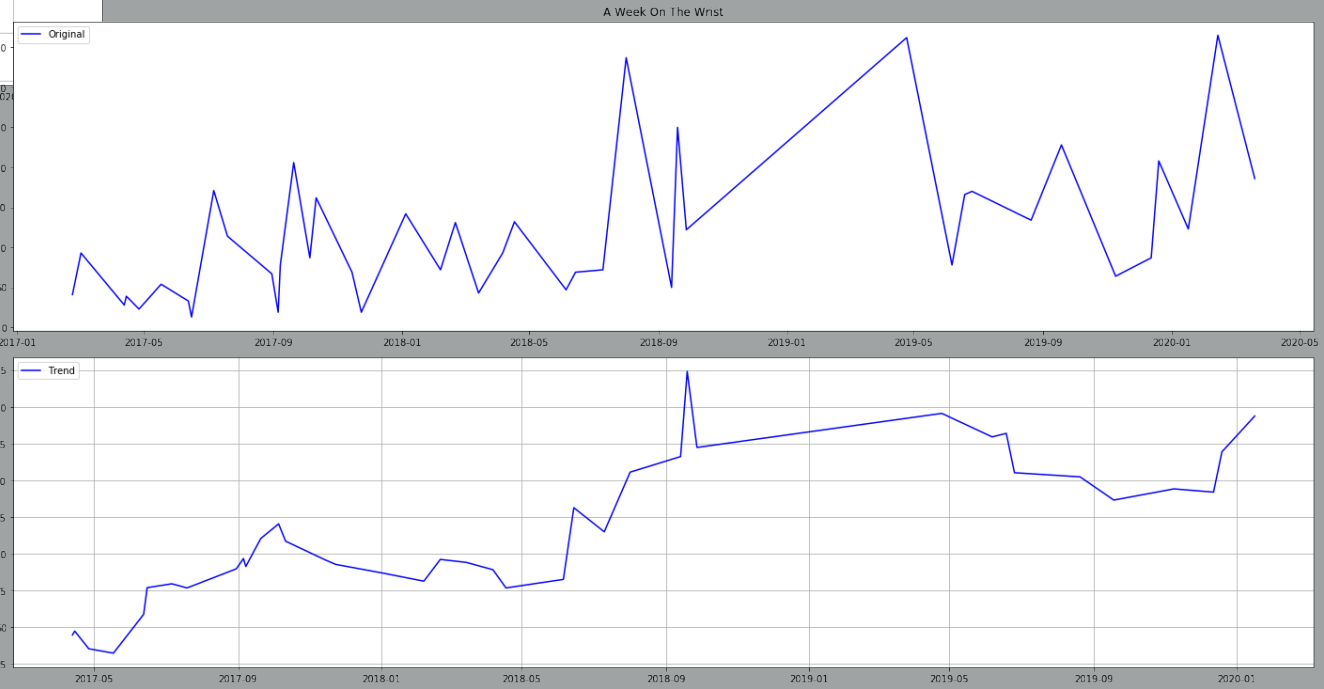


Categories with growing engagement



In-Depth

A Week On The Wrist



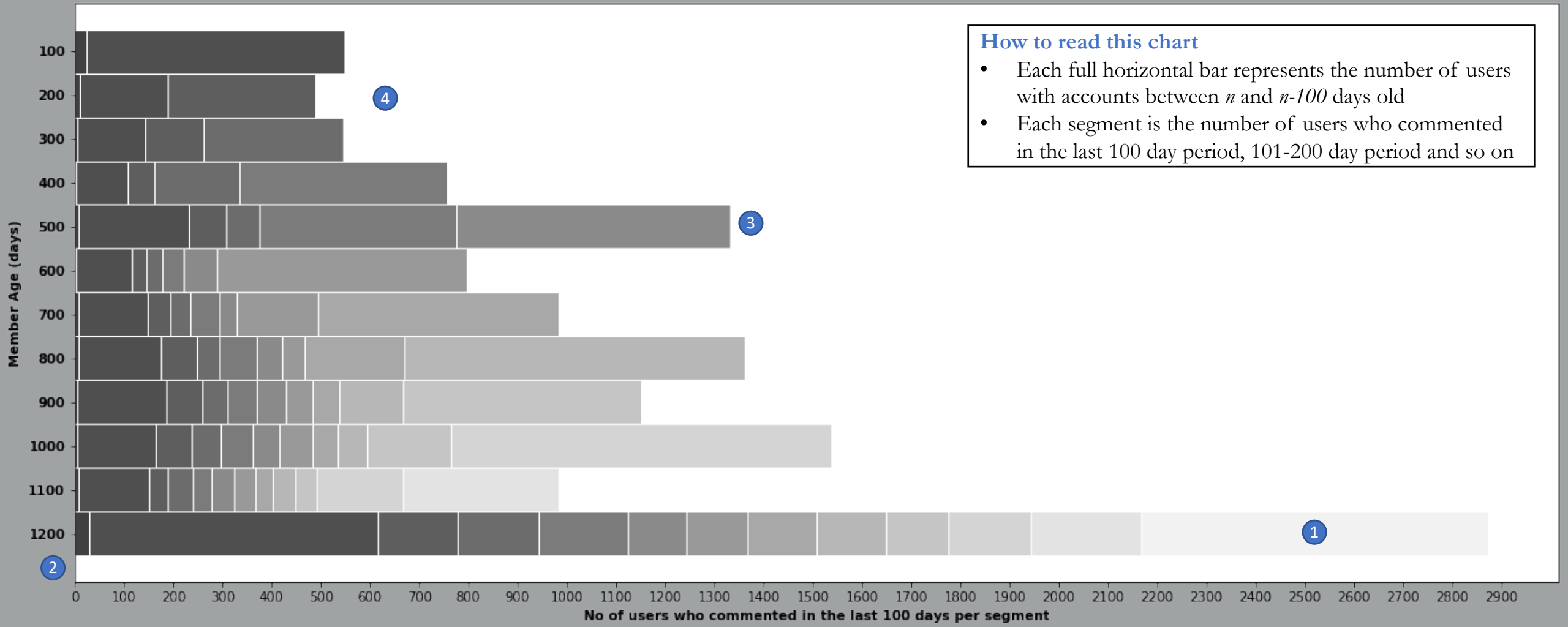
Most prolific commenters (normalized by account age)

Rank	Username	Prolificness Score	Comment Count	Account Age (days)	No. of Watches in Profile
1	Shatners	1.38	721	521	0
2	ICH	1.38	1392	1009	1
3	Gav	1.35	499	370	5
4	GreatScot	1.34	127	95	0
5	JackForster	1.11	1324	1190	0
6	Bside	0.98	1171	1190	0
7	Boman	0.97	625	644	5
8	wkf	0.90	1072	1190	12
9	PaulMiller	0.85	1008	1190	0
10	ripwatch	0.81	720	886	0
11	Oliver_H	0.77	27	35	5
12	Yev	0.77	703	917	5
13	CynicalBastard	0.76	626	825	0
14	TheOmegaMan	0.75	853	1131	10
15	Putito	0.73	136	187	1
16	AJ117	0.72	244	340	0
17	ThicknessMatters	0.71	762	1070	0
	…	…	…	…	…
20	Cole	0.63	271	429	9
	…	…	…	…	…
82	BenClymer	0.27	318	1190	3
	…	…	…	…	…
124	CaraBarrett	0.22	264	1190	2
	…	…	…	…	…
133	ghariyaan	0.21	131	613	7

Prolificness is calculated as $\frac{Comment\ Count}{Account\ Age\ in\ days}$

Included some Hodinkee Staff and myself for reference
(I'm surprised and disappointed my score isn't higher ☹)

User age and engagement pyramid



1. 700 of the oldest users on the site haven't commented in 1,200 days
2. There is a continuously engaged cohort of users across all account ages
3. Users picked up around December 2018 continue to be more active
4. Fewer new users have signed up in the last year than in the previous two years

Key Takeaways: Most users who signed up didn't remain engaged and fewer users have been signing up over time

Future explorations

Comments

- Proportion of comments posted in reply to other comments
- Longest conversations (most comments in reply to the same top-level parent comment)
- Most liked comments
- Likelihood of comments being left on older articles
- Comments distribution by brand

Users

- Watch collection analysis (number of watches, types, etc)
- Favorite brands
- User location distribution
- Online profile

Random dead ends

- Time to First Comment after article is published: ‘datePublished’ in the HTML for the page is not the time the article is actually visible to users
- Member locations: difficult to analyze since this field is free-form
- Tags-based analysis: highly skewed assignment on articles, mixed tag types (article categories, events, watch types, watch brands, etc)