# [MKP] - Hypothesis, Annotations, and Engagement

> An Exploration of public annotiations made during the #publicknowledge course during the Fall term
> 19' at SFU

Author: Asura Enkhbayar Date: 05.12.2018

**Table of Content**

## Materials

- Full Jupyter notebook including code for analysis and plots: notebook
- Presentation used for the last class of the course: link
- Detailed survey results: spreadsheet
- About Hypothesis: homepage

## The Data Sources

The data used for this analysis comes from three different sources.

### 1. Making Knowledge Public - Course Material

The course structure and outline as described on the syllabus has been manually saved in spreadsheets.

### 2. Survey among students

A survey among students was used to (1) identify and confirm Hypothesis usernames needed to differentiate
comments relevant to the course and annotations created by other public users.

### 3. Our annotations retrieved from Hypothesis API

Hypothesis provides an API to programmatically retrieve annotation data for individual URLs. The URLs to
the readings were available in the course material. The API returns the total number of annotations for each
URL and a list of annotation object. Each annotation contains among other things:

- A unique ID for this annotation
- Date created
- Date updated
- The full text of the comment
- References (other annotations that this object was a reply to)
- The username of the author

## The Data

In total, the course consisted of 12 weeks and 28 annotated readings. Hypothesis found 1431 annotations for these 28 URLs, while 24 of these were created before the first class of the PDC. Of the 1417 annotations that were created during the 12 weeks, 1286 (91%) were created by students of this course.

9 students filled out the survey, while 7 provided a complete set of answers.

> **1286 annotations** created by **14 students** (and Juan) during **12 weeks** of classes on **28 readings**.

## What can we learn from the data?

Given that the data from Hypothesis is available, I decided to use this opportunity to explore the notions of annotations & engagement within the context of the President's Dream Colloquium.

In the following section I will present a few initial insights and attempts to think about the meaning of annotating papers and the implications of the availability of knowledge about annotations as structured data.

### When did students annotate?

Figure 1 visualises the unsurprising fact that human beings (and students) work towards deadlines. The size of bubbles representing new comments on the readings increase in their size once the weekly lectures (red lines) approach.
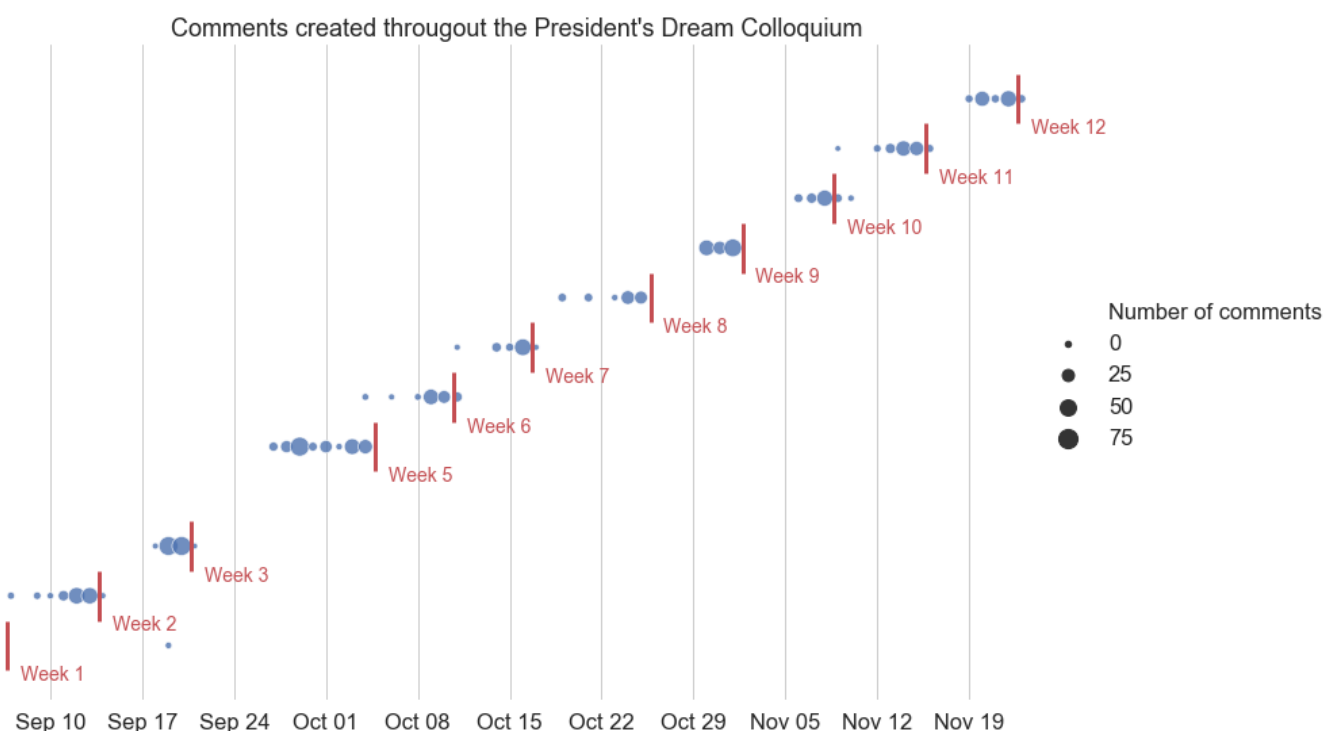
*Fig.1: Overview of comments during the PDC.*

We can depict this circumstance in an even more convincing way if we look at the days until the relevant class for each annotation. We would expect to see majority of the activity within the week before each class, especially peaking just before the final day. This is exactly what we observe in figure 2, which shows that more than 65% of all annotations are created on the same day as the class or the day before.
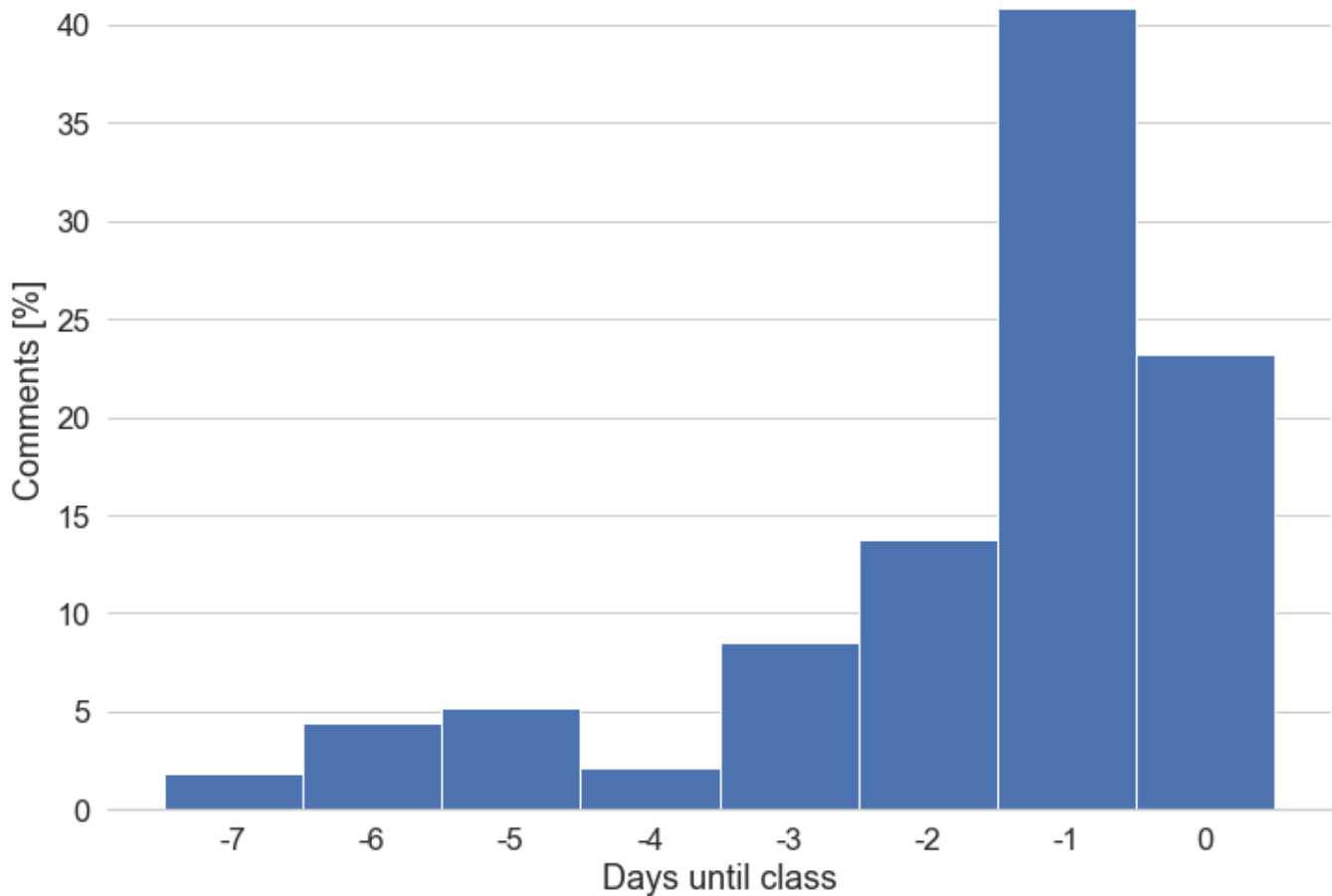


*Fig.2: Number of days before a class took place*

These findings are not entirely surprising even though I was curious if the data would show some annotations that were created outside of the expected margin of plus/minus a few days of the lecture.

## How did students annotate?

Now that we know when students annotate their readings, we can start to look at which readings were annotated more often than others. Figure 3 shows that the number of comments fluctuated around 100 comments per week with the exception of week 5 (Luke Terra's lecture on *University-Community Connections*). At this point, it is interesting to note that that week also received the most votes for the *most engaging* week (3 votes ahead of Robin DeRosa with 2 votes).
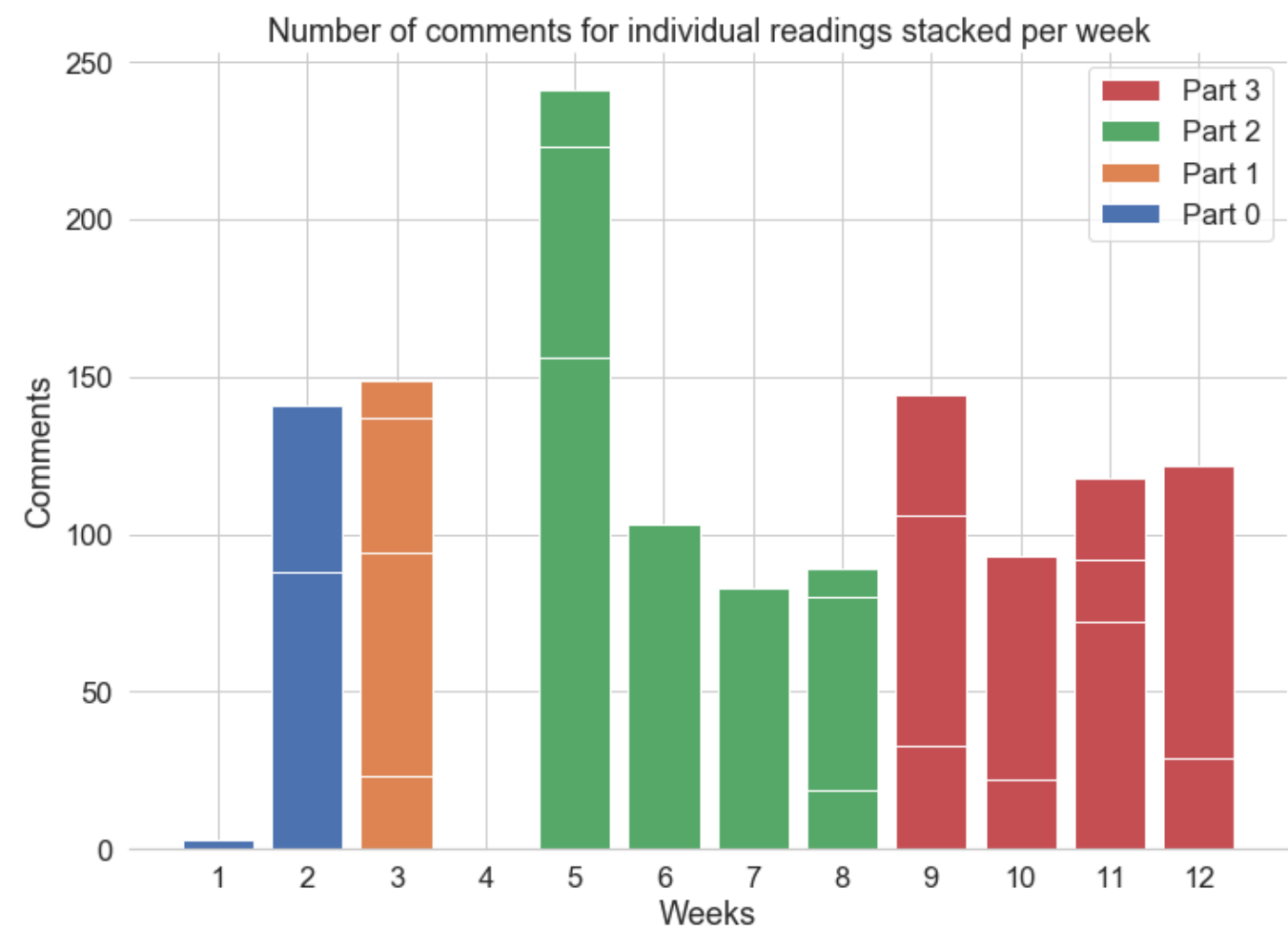
Fig.3: Each stacked sub-bar represents one reading. Multiple readings in the same week are stacked to represent the total of the week. Colors are according to the assigned part in the syllabus.

But does it make sense to simply look at the number of comments created on the document to evaluate/measure/predict [*insert vague concept of choice*][1]? My simple assumption was, that writing longer comments should take you longer, thus, leading to a lack of time to create more comments. More comments -> shorter comments. Let us question the data:
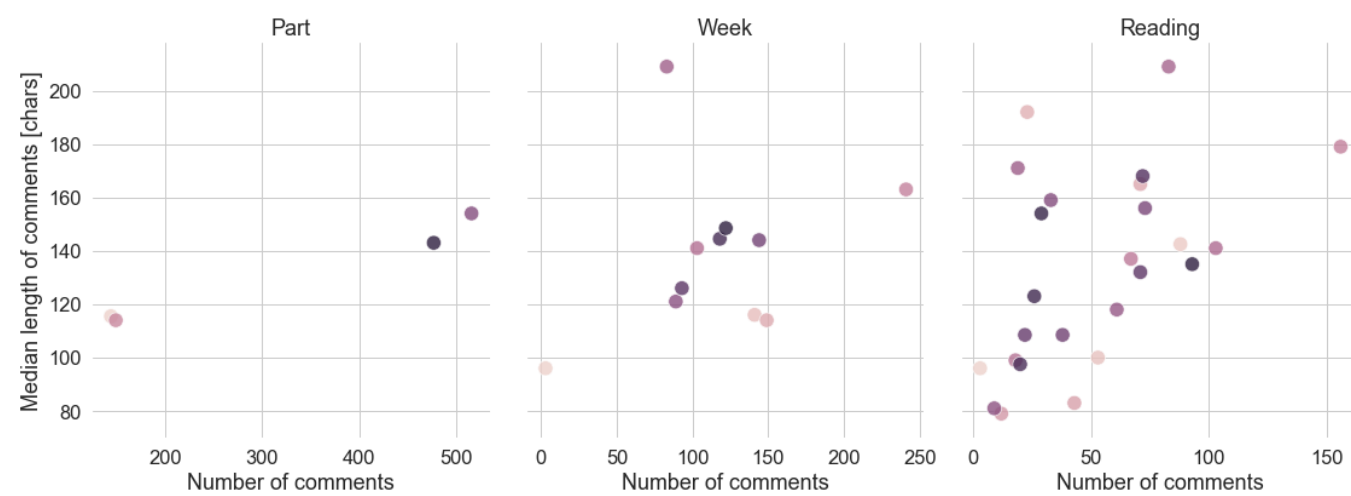


Fig.4: Comparing the median length of comments with the number of comments for each part, week, and reading.

As we can see, my highly sophisticated hypothesis does not hold in the face of the equally highly sophisticated data. I am refusing to offer any correlations at this point, as I do not think that the reduction to a number would improve our insights, and will simply ask you to have a look at the scatterplots in figure 4 which show that the number of comments and the median length seem to show a mild positive correlation for aggregations across the parts, weeks, or single readings.

But what about comment lengths and comment frequency for individual users? Can we model different *annotation styles* based on these two variables?
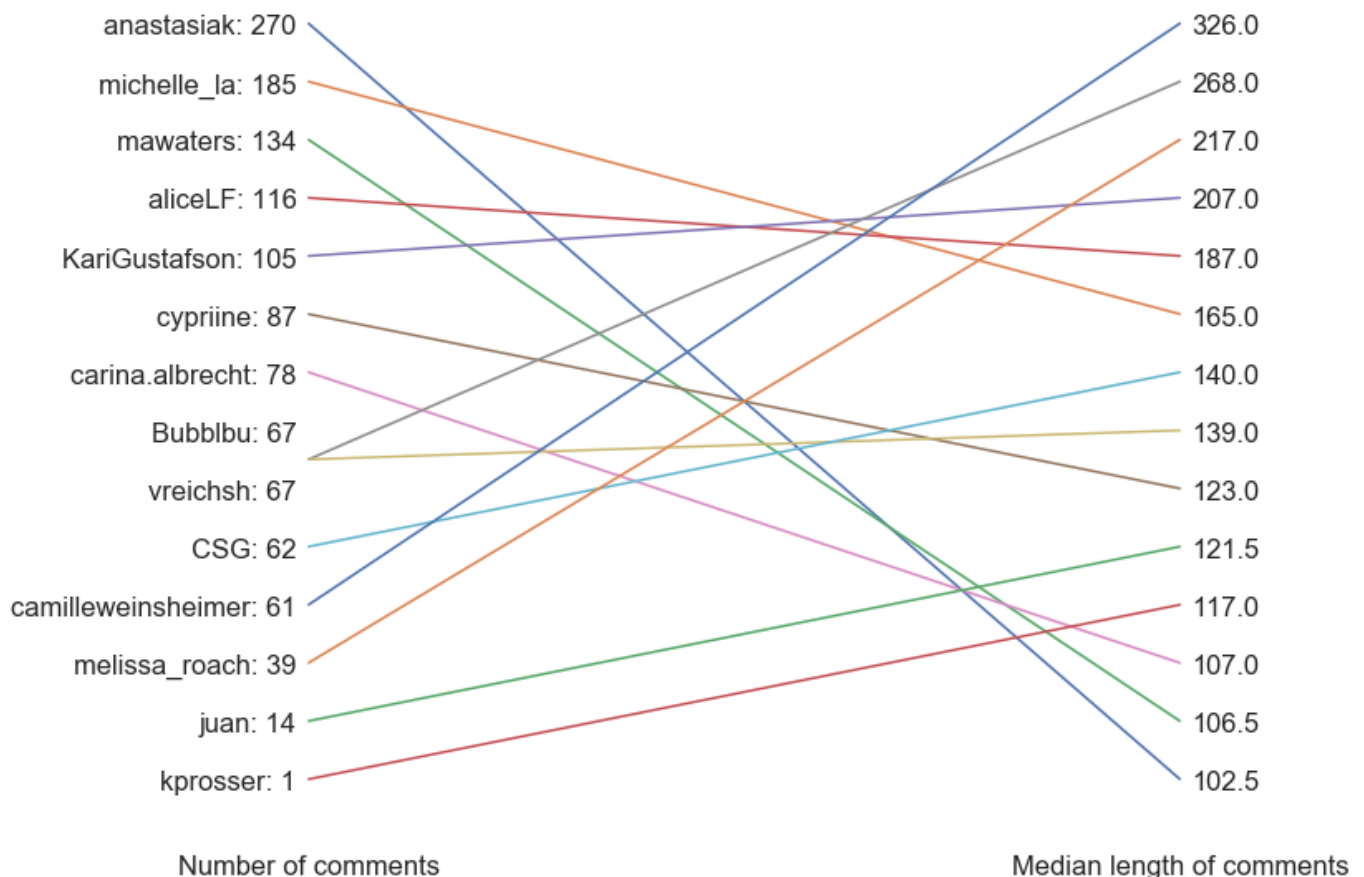


*Fig.5: The changes of user rankings according to the number of comments created and the median length of comments*

Interestingly, for our limited dataset, there seem to be a few different types of annotaters in our class.

- *High frequency, short texts*: anastasiak, mawaters
- *Low frequency, long texts*: Bubblbu, camilleweinsheimer, melissa_roach
- *Mid-high frequency, mid-long texts*: aliceLF, KariGustafson, michelle_la
- *Mid-low frequency, mid-short texts*: carina.albrecht, cypriine, CSG, vreichsh

Juan was excluded as the course instructore and kprosser due to a single comment.

**Some other insights...**

Just a few extras to show off other possible analyses based on the available data.

Figure 6 shows how the creation and update times contained in the annotations could be used to (very) roughly estimate the reading time for an article. Some obvious problems are work sessions spread across

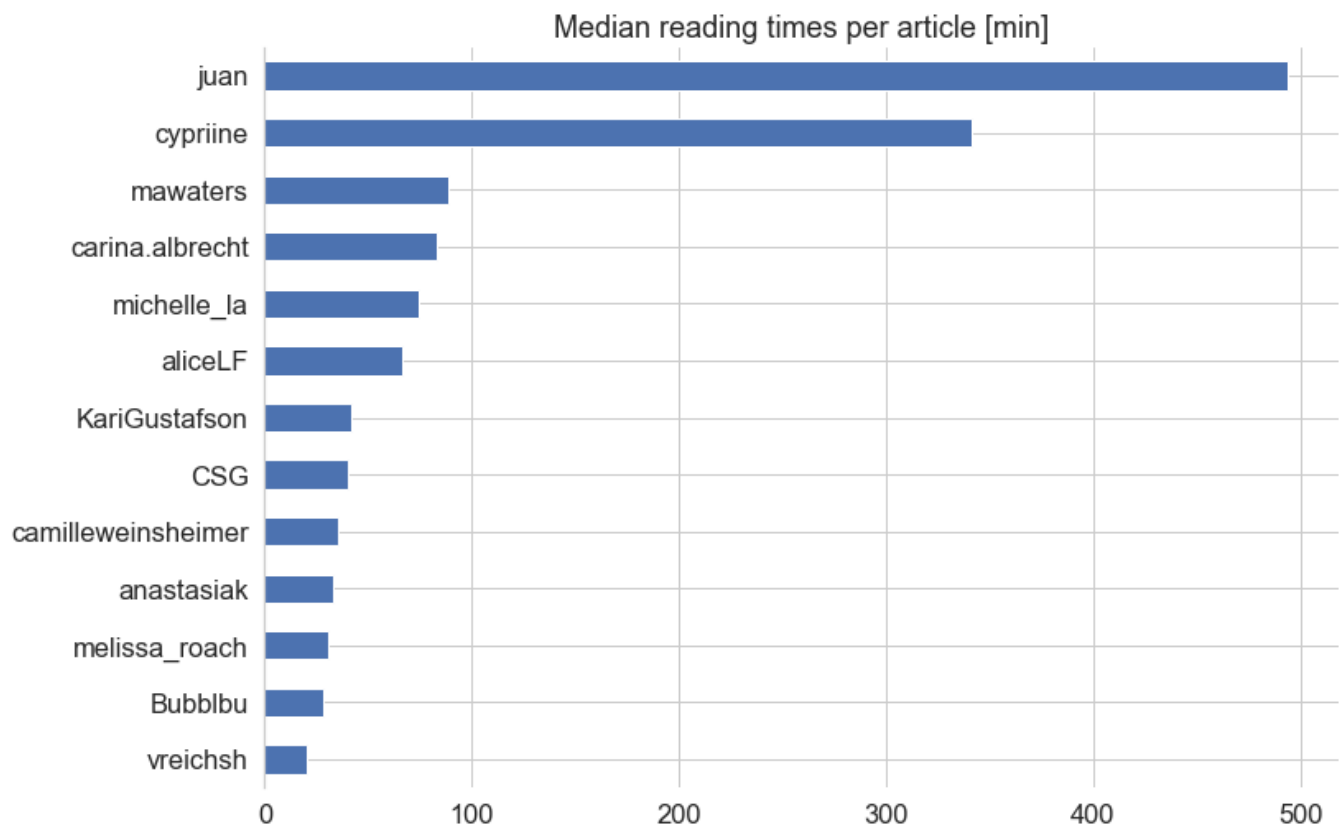long timespans and that the actual activity of reading and thinking does not necessarily occur linearly.



*Fig.6: Median reading times per article*

While reading comments I thought that linking to other resources (e.g., Wikipedia for definitions, follow-up research, or relevant news coverage) represented a special kind of annotation that felt more *engaging*. I was curious to see how frequent these comments with links to resources were in the data. Turns out that only 106 (8.2%) of all comments contained a URL. Figure 6 shows a breakdown of the most avid linkers among our group.
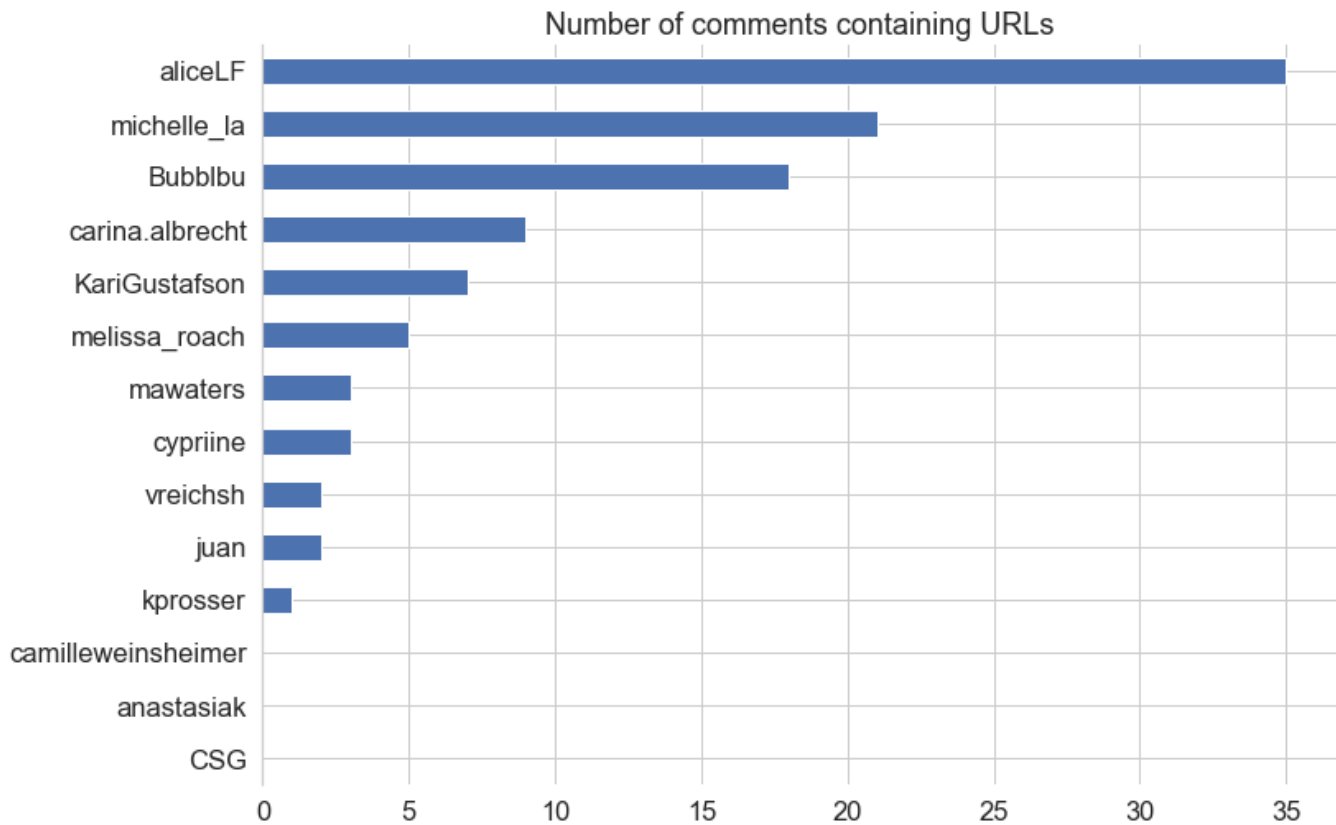
*Fig.6: Number of comments containing URLs*

## Predicting engagement with annotations

Finally, I wanted to present a very quick and dirty attempt to quantify the predictability of *engagement* of the parts, weeks, and readings in the syllabus. The survey results for these three categories were used as target variables which were to be predicted by the number of comments and median length of comments for individual users.

The, obviously naive, hypothesis is that *highly engaging* readings (or weeks or parts) will evoke more or respectively longer annotations.

The following graph shows the difference in the rank between the user's actual *most engaging* paper/week/part as reported in the survey and rank of the same reported paper/week/part as measured by number/length of comments. This means, that ideally we would see two matrices of 1s meaning that each self-reported item also took the 1st place in the measured ranking.

## Predicting the "most engaging" paper by ...

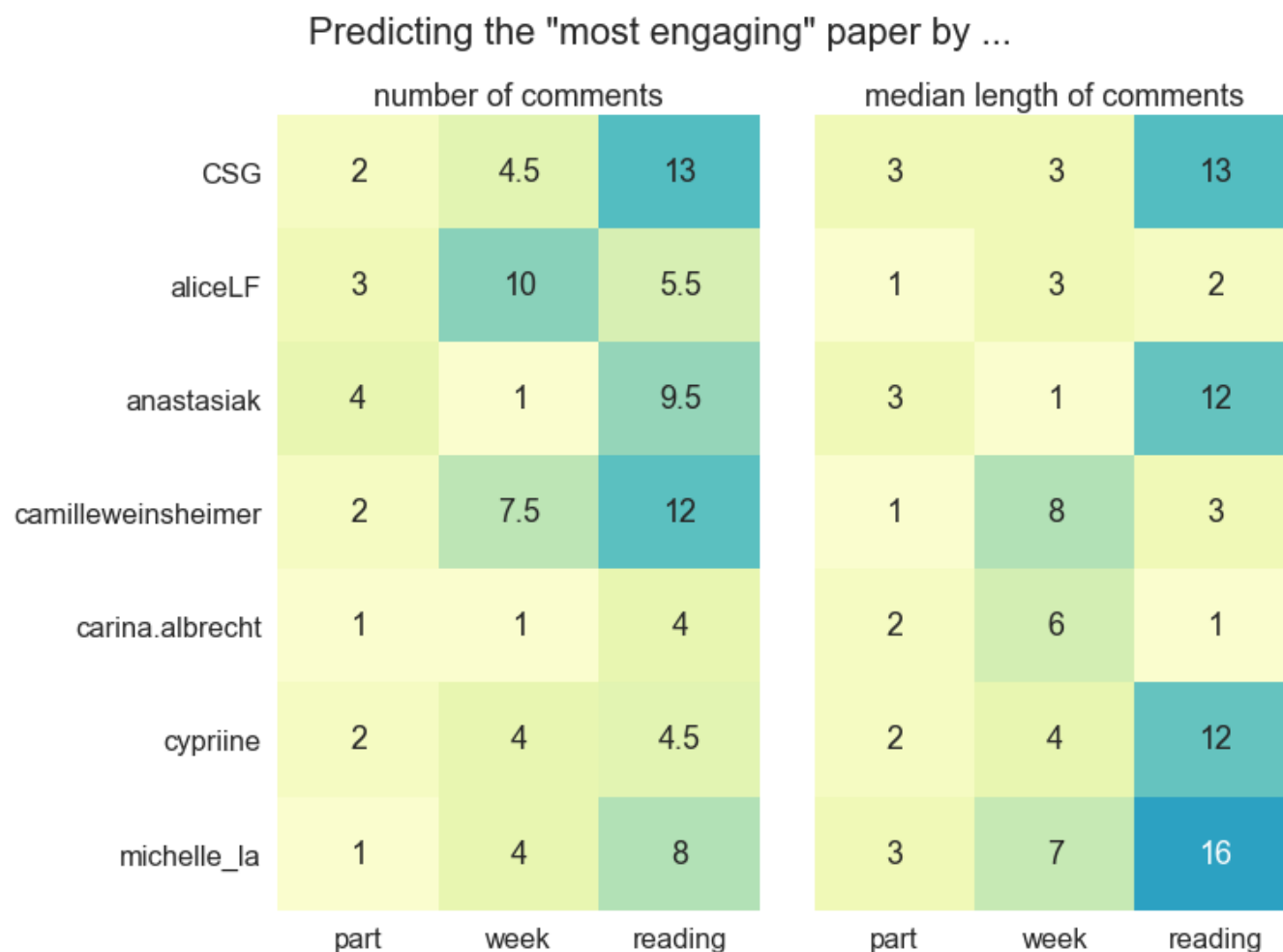| | number of comments | | | median length of comments | | |
|---|---|---|---|---|---|---|
| CSG | 2 | 4.5 | 13 | 3 | 3 | 13 |
| aliceLF | 3 | 10 | 5.5 | 1 | 3 | 2 |
| anastasiak | 4 | 1 | 9.5 | 3 | 1 | 12 |
| camilleweinsheimer | 2 | 7.5 | 12 | 1 | 8 | 3 |
| carina.albrecht | 1 | 1 | 4 | 2 | 6 | 1 |
| cypriine | 2 | 4 | 4.5 | 2 | 4 | 12 |
| michelle_la | 1 | 4 | 8 | 3 | 7 | 16 |
| | part | week | reading | part | week | reading |

*Fig.7: Differences between the rank of self-reported most-engaging items (rank: 1) and the ranks of items as measured by number and length of comments.*

## Contact

I would love to hear about your thoughts and other ideas for further analyses. Feel free to submit a new issue or simply shoot me a text at asura.enkhbayar@gmail.com.

## License

---

[1]: In this wonderful case, I have been using the term *engagement* to refer to an ill-defined concept that will unavoidably evoke other concepts such as interesting, critical, questionable for each one of us. Knowingly, I am playing along, and trying to see where the numbers take us.