Introduction

Many goals in NLP focus on analyzing a chunk of text, and making some sort of prediction or identification about it. In the CS6742 course, we were given two articles [GK10] and [ana09] to read, and come up with interesting research proposals.

[ana09] acknowledges the failure of traditional moderation methods in ensuring productive conversation on forum posts and threads. The article explores changes to traditional moderation systems, and puts forth the idea of evaluating conversations in threads as an indicator of engagement within the online community, instead of just looking at comments.

While several interesting research questions were proposed, we will only look at the following problem.

Research problem

The problem is stated as such: *Can we predict the dynamics of any two active users, based on their previous posts?*

Our analysis will be done using data from BC3's blog corpus, in particular the Slashdot blog conversations. Using this data, we would like to answer the following questions:

1. We would like to know if we can predict the trend of a thread, given the earliest k comments in the article. The trend here can be the length, a moderation class, the moderation score, or some derived feature based on the text such as aggressiveness.

2. Given the probability vector $\mathbf{p}_A = (p_{1A}, p_{2A}, \ldots, p_{nA})$ for any user A to create posts of quality $1, 2, \ldots, n$, we would like to estimate the conditional probability vector $\mathbf{p}_{A|\mathcal{E}}$ of user A, given that some event $\mathcal{E}$ has occurred. Here, $\mathcal{E}$ could be a shift in topic of the thread, or the entry of a new user.

Proposed techniques and processing and selection of data

1. We begin addressing research question 1. by analyzing the comment lengths and moderation scores in a thread. For this preliminary analysis, we represent a thread as a collection of all comments sorted chronologically. An alternative could be to represent a single thread as a path from root to leaf.

2.

-

-

[BKLDNM13] also proposes a metric based on whether two users are connected to each other[1]. Although there is nothing similar to that for the Slashdot data, we propose a metric of the following: We take a look at the datestamps of the posts of any two particular users, and construct an interval around each datestamp, and assume that this is a user's available time online. We take the ratio of the intersection of two users' available time to their total time to be a metric[2].

We hope that we can use these metrics to eventually classify the rest of the `None` posts. Then, we can then look at the empirical probability and conditional probability, and do more analysis of the data.

Ideally, we would like to use the entire Slashdot blog corpus and split it into a testing and training set, but it is highly possible that some comments will be hard to classify, and we might have to discard them.

Preliminary results

While the Slashdot data has comments with a moderator score and a category, exploratory data analysis shows that the majority of categories are labelled as `None`. However, we analyze preliminary results to see if there are any trends in moderation score.

The breakdown of the classification of comments are as follows:

```
{'Troll': 14932, 'Funny': 40672, 'None': 464104, 'Flamebait': 7456,
'Redundant': 4792, 'Offtopic': 11384, 'Informativ': 40188, 'Interestin':
50168, 'Insightful': 73864}
```

which really necessitates looking at the `None:` posts and classifying them.

One interesting preliminary result however, came from the metric of post length, and we look at the top 25 posters. We made plots of each user's post length and the conditional distribution of their post length when another user is in the same thread. There are several examples of these plots on subsequent pages.

---

[1]This study looks at Facebook and Wikipedia editors. Facebook users are connected if they are mutual friends, and Wikipedia editors are connected if one an editor posted on another editor's page.

[2]For example, if user a has the time intervals $T_1 \cup T_2 \cup \ldots \cup T_n = A$, and user b has the time intervals $T'_1, \cup T'_2 \cup \ldots \cup T'_m = B$, we would look at the ratio $\frac{A \cap B}{A \cup B}$.
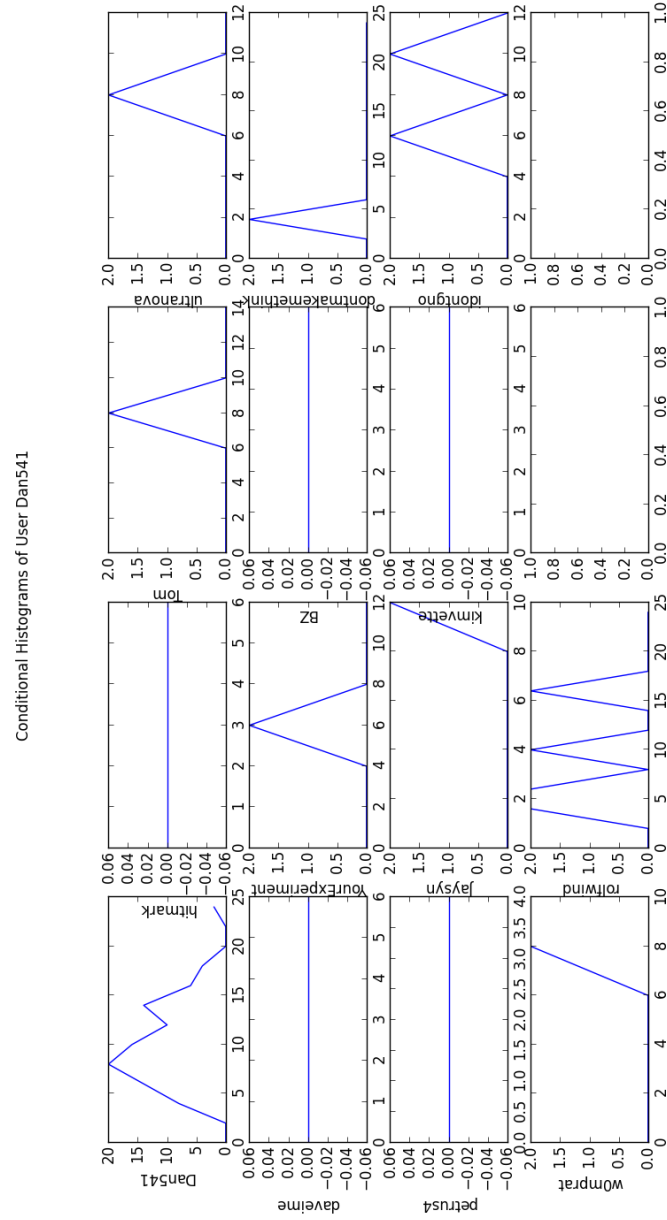
Figure 1: Joint Distribution Plots for user `Dan541` and Other Active User

Here, the x axis denotes the square root of the number of characters in a post, and the y axis denotes the number of threads where that occurs.

What we learnt

Based on the conditional/joint distributions, we can see that users make differing lengths of posts depending on which particular user is in the conversation. However, this is for a small subset, and we want to see if this has any significance in a bigger set of data.

Roles played

Keegan - Wrote report with Ben.
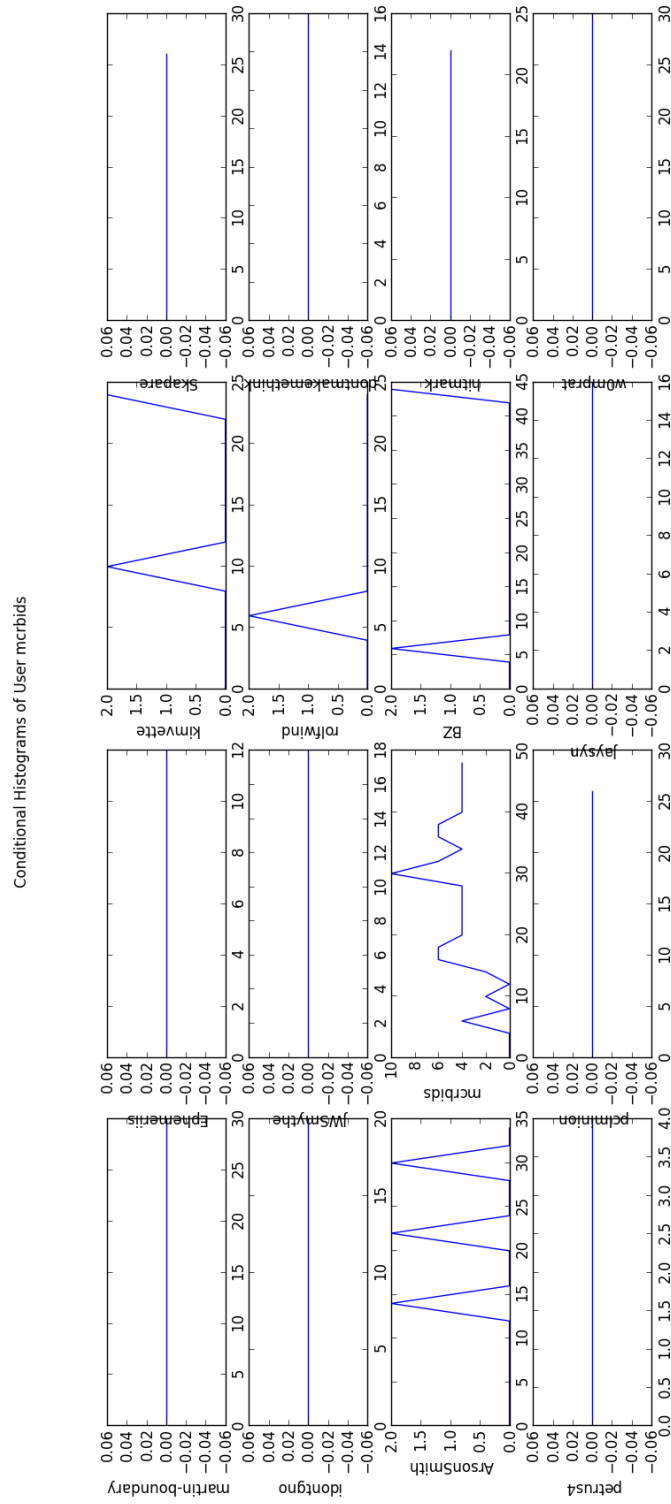Ben - Super Duper Code expert and Python debugger. Wrote report with Keegan.

Figure 2: Conditional Plots for user `mcrbids`

References

[ana09]   anaesthetica.   Attacked   from   within.   http://aiweb.techfak.uni-
               bielefeld.de/content/bworld-robot-control-software/, Mar 2009.

[BKLDNM13] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-
               Mizil. Characterizing and curating conversation threads: Expansion, focus,
               volume, re-entry. In Proceedings of WSDM, pages 13–22, 2013.

[GK10]   Eric Gilbert and Karrie Karahalios.   Understanding deja reviewers.   In
               Kori Inkpen Quinn, Carl Gutwin, and John C. Tang, editors, CSCW, 2010.

[Ott09]   Jahna Otterbacher. 'helpfulness' in online communities: A measure of mes-
               sage quality. In Proceedings of the SIGCHI Conference on Human Factors in
               Computing Systems, pages 955–964, New York, NY, USA, 2009. ACM.