11719 Assignment 2

Seungwhan (Shane) Moon, Saloni Potdar, Lara Martin

March 2, 2014

1 Observations

This section provides our qualitative analysis of the MOOC Python Course dataset.

1.1 "Student Leaders"

Our main focus in our analysis is that of "student leaders" and their position in the MOOC threads. One of the observations that we have made while digging through this corpus is that there are certain students that rise to a position that is similar to an instructor or TA. Our goal is to find out how these students mark their leadership and how the other students in the course respond to them. A way to test this is to compare the student leaders to the actual instructors and teaching assistants of the course. If the students are accommodating to the way the instructors/TAs are speaking, then they are assuming the role of "student leader."

Here are some examples of posts from TAs:

> **TA**: I confess that I have no idea why that is happening. Perhaps you can provide more details regarding your system & python version to help identify what's going on there.
>
> **TA**: Yes, spacing is extremely important. Not because your code will cease to function if you don't space correctly, but because it increases readability, so it will help other people read your code, and it will help you read your code when you come back after a vacation or whatever.

Here are examples of posts from instructors:

> **Instructor**: We haven't made big changes to the course, although we are cleaning up the videos as we go – improving the sound, making the videos a bit smoother. We also proofread and updated the exercises.
>
> **Instructor**: Thousands of people completed the first offering of this course without a textbook. It really isn't necessary!If you want a free book, there are several listed on the Course Wiki (the link is in the navigation bar to the left).

Here are examples of posts from students who speak similarly to leaders:

> **S1**: The spaces between the values and the operators are not required. The previous answer which indicated that an additional blank line is required after the last line of a function to terminate the block of code is the correct answer.
>
> **S2**: Hi,Active has a Python package that is free for educational purposes. I've looked into it, but still have my doubts concerning installing another Tk framework in parallel to the one supplied by Apple. You might want to take a look at it; a friend of mine who does a lot of Python development

uses this.Hope this helps.

Here are examples of posts from students who do not speak similarly to leaders:

S1: i have the same problem, and i press enter. I don't know what i'm doing wrong.P.S.: Ok i have the answer. Press enter two times after writing de return part.

S2: Yay :D I am glad someone helped you out. Text

It appears that the instructors talk about administrative business for the course and use a very formal, but sometimes friendly, tone. We believe that if we are going to find out who is a student leader, it would be best to compare them to the speech of a TA, since they speak more similarly. This is probably since the TAs are technically students who are in charge of answering questions. Non-leader students speak far more informally, as if everyone was their peer.

1.2 Correlation between the number of votes and influence

One of the tags that we have found in our Python corpus is that each post includes a number of votes that other users have given it. We hypothesize that this number should correlate, to a certain extent, the amount of leadership the author of the quote has. After looking at the data, we have found that the number of votes that a post gets in a thread does seem to point to leadership and not simply the entertainment value. Furthermore, most of the highly rated posts are authored by students and not instructors or TAs.

The post with the most votes throughout the entire corpus, at 130 votes, is actually by a student who is giving advice on completing a homework. The person is clearly some sort of "student leader" and gives to-the-point advice with a bit of formality. The next most popular post (82 votes) was a comment and a link to a PDF version of a lecture summary that was already online as a webpage. From then on, the number of votes drops to 56, and then the other posts continue with a more gradual decrease in the number of votes. The post which has 56 votes is one in which a student helps straighten out some definitions. Interestingly, neither a TA nor an instructor shows up until the 18^{th} most voted post. In fact, only 13 out of the top-100 voted posts were written by an instructor or TA. We can see no correlation between the number of votes and how long the entry is. Some high-rated entries are just links, while others are entire paragraphs. Other comments that are still in the top 100 most rated include funny comments, such as this one (at 14 votes), which is talking about the final exam:

> Speaking as the non-elected representative of the chaff, "Ouch."

Due to these discoveries, it is reassuring that we could possibly use the number of votes as a check during our analysis of student leaders, and that perhaps, a high number of votes on a particular post could imply that it was written by a student leader.

1.3 Time-varying social interactions

We have observed that the MOOC discussion forum exhibits a unique developmental process. The most notable feature is that students come in and leave the forum in waves throughout the course. We can thus group the students into cohorts depending on the order of student's first appearance on the forum, and study the interactions among the different cohorts of students. Figure 1 shows the student cohorts and their activities on the discussion forum throughout the course.

In Figure 1, it can be seen that the first cohort (blue) continues to actively participate in the discussion throughout the entire course. The social interactions among the students would almost certainly be different for each time frame, as new students come in and out constantly.

This observation allows us to think of several approaches which we can take to analyze the dataset. In Section refsec:leaders, we observed that there are student leaders who voluntarily help other students by answering their questions. According to the definition of a power difference proposed in [1], student leaders

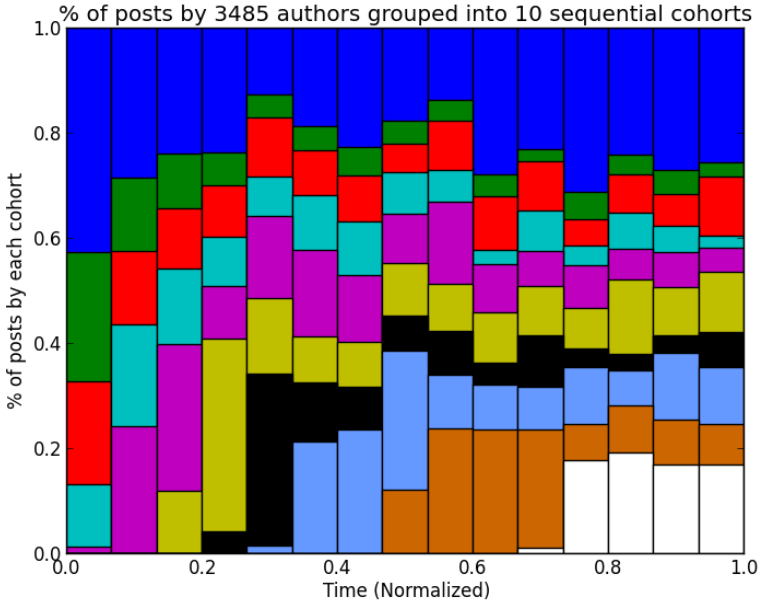

Figure 1: Blog posts distribution of 10 cohorts throughout the course. Each color represents a cohort, where the blue bar on the top represents the cohort of the first 10% of students, and the white bar on the bottom represents the cohort of the last 10% of students. X-axis denotes the normalized timestamps of the post, and thus each column represents a timeframe at which we aggregate the activities by different cohorts. The posts by instructors or TAs were not included in this plot.

can be considered as having higher powers than other students, where the power difference arises through dependence (students need answers to their questions from student leaders). Therefore, by grouping the posts into their authors' cohorts and the corresponding time frames, we can study how the influence by student leaders is accumulated at each time frame, and thus propagated to the following cohorts of students.

There are several relevant works that may help analyzing this observation. [2], for example, claims that linguistic change captures the relation between a user and its community. While this paper studies the lexical innovations within an online community and users' adaptation to the new community norms over a relatively long time (10 years), we do not believe that we will be able to detect changes in the community norms due to the short lifespan of the course. However, we would like to study whether the language accommodation is cumulative, and whether its influence can be propagated gradually among the group of people who were in direct communication.

[3] provides a great framework for measuring influence propagation on a social network and discovering leaders from community actions. However, applying this methodology to our dataset will require extensive modifications primarily because the community action (language accommodation) is much more subtle and thus hard to quantify.

1.4 Discourse structure - who is replying to whom

The challenge in analyzing the posts on a thread-based discussion forum is to identify who is interacting with or replying to whom. In this section, we provide examples for several different ways to detect how students are interacting.

1. **Addressing names** : People might address people directly in a conversation with their names. In discussion forums this is a very valuable piece of information as we come to know who is talking to

whom when there are a large number of people in the conversation. This will help us to easily model the influence propagation and also validate whether we have identified the leaders correctly (by looking at how many times people have mentioned them).

> **jonathan lung**: Yes! The only text files I don't edit on a Mac with TextWrangler are LaTeX files and a few files I edit through vi/nano (and I edit a lot of different kinds of text files).
>
> **Steven Johnson**: Thanks Jonathan

2. **Redundant posts - people ignore previous posts** : People in the forum might ignore or not read the previous posts before they reply. This may lead to redundant posts, which makes it difficult to analyze the conversations. To identify influence in this case we may need to adopt other techniques, since if we consider a 2 message window, it may seem like the second person is fully influenced by the first person.

> **A**: The order of precedence in python is pretty more like C++. Also want to say thanks to Jennifer Campbell and Paul Gries for this opportunity.
>
> **B**: Pretty much most programming language has the order of precedence of C++.

3. **People read all the posts above and make a new contribution** People in the forum may read all the previous posts before they reply. This is a normal behavior we expect.

> **A**: Is there a way that I can change quality of the videos? I don't want to download them but for my internet connection video quality is too high. ···
>
> **B**: Very strange the sub are not in sync.Have you already tried another browser?Or just the options on the right-bottom corner of the videos?!Try this and tell us if you resolve...
>
> **C**: The video quality button is not visible to me too. I think their goal is to offer ···
>
> **B**: You cannot change the quality (AFAIK), but in some cases you can change ···
>
> **A**: So in Opera subtitles weren't displayed but in chrome there were also too slow either in vlc ···

> **A**: Remember how area is calculated. There is a division in the return statement which produces a float. So you get a float as a result.
>
> **B**: "Well, the operator used is the floating point division (""/""), In order to return an integer, we need to use the integer division (""//"")
>
> **C**: Thanks! Got it.
>
> **D**: or you can change the output in the return. for an example if you use:return int(4 / 2)you will get 2 instead of 2.0
>
> **A**: @ C area computation of a triangle requires a float resultant because of the division involve in it. ···

4. **Quotings** : People might quote others in the conversation. On one hand this might show who the speaker is responding to, but it may again qrongly show influence. Thus we need to get rid of these quotes in the conversation.

> **A**: hello everyone. i just wanted to know the answer to question#8 .consider this code :
>
> ```
> x = 3
> y = 5
> x = y
> ```

After the code above has executed, what value does y refer?

B: this was my answer:Consider this code:

```
x = 3
y = 5
x = y
```

After the code above has executed, what value does y refer?id (x) = id (y), y refers to the same memory location ···

A: "We have intentionally left out tests involving time zones that are not on the hour: ···"Are they saying that they only passed the time in hours and we need to handle cases if they were passed in minutes and/or hours?

B: "We have intentionally left out tests involving time zones that are not on the hour: ···" Can any one explain this please? I saw someone else asked the same question but I couldn't understand ···

5. **Anonymous posts** : We observed that this will be difficult to model as we don't know who exactly is talking. There may be number of people in the forum who may comment on posts as anonymous. Trying to label and treat the anonymous people as actual participants is a challenge we will face.

Frank Sullivan: just look at the available ones

Richard Jonathan Huijgen: Maybe we can all add our location to the map Claudia has created? :)

Anonymous: 只有一个中国人吗?

Anonymous: Frank, I'm also taking this to further my crypto studies :)

Anonymous: When you DEFINE a function the function definition has parameters, when you CALL a function the call has arguments.

A: This is the best answer.

B: Thank you, Anonymous! so when you CALL a function it is always a built-in function, right?

B: Anonymous, I am not so sure.....Take the built-in function ···

Identifying the discourse structure is extremely important in modeling the computational model. [1], for example, models the language accommodation between a speaker and a target locally, by assuming that a target accommodates to the preceeding utterance that a speaker has made. However, as seen in the examples above, we need a better way to process a thread-based discussion forum like our dataset. We can either implement a communication pair detector that can deal with every case we have identified, or make a more global and general assumption (e.g. a new post on a thread is influenced by all of the proceeding posts on the same thread).

1.5 Off-topic replies

There were a few off-topic replies in the forum. These might be the result of being unfamiliar and new with the forum interface, or it may be because the students want to draw attention.

So the main question is how far to look for the dependency amongst posts for influence? We think a good approach will be to look at first few posts of the discussion and a few just before the post being considered. We think this is a reasonable approach as these posts are most likely to be looked at before responding.

1.6 Code snippets

We have observed that students frequently use code snippets in their comments. Code snippets may appear in a new text block or inline within the comment.

- Example 1.

> hello I am tring the following code ,but it is giving 0 every time I run this .
>
> ```
> nucleotide = 0
>  for char in dna:
>      if char == nucleotide :
>          nucleotide = nucleotide + 1
>  return nucleotide
> ```
>
> Can someone please tell me what is wrong here ?

- Example 2.

> let me fix the mistake your if statement need to check to see `if char in "ATGC"`: sorry I'm getting sleepy, my error

We believe that these code snippets will be mostly distractive in our text analysis. One solution would be to build a parser to detect a code snippet and assign a tag instead to each block of code. Example 2, however, shows potential challenges in differentiating code from the given text, especially because the Python programming language is designed to resemble our natural language.

References

[1] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, "Echoes of power: Language effects and power differences in social interaction," *Proceedings of WWW 2012*, 2012.

[2] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, "No country for old members: User lifecycle and linguistic change in online communities," in *Proceedings of WWW*, 2013.

[3] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Discovering leaders from community actions," *CIKM '08*, 2008.