Successful Threads Memo 2017-12-29 – Cultural Similarity and Presentation Questions

LIWC distance to execs

Description of Appraoch

I used the code provided by Amir's team to reproduce the calculation of a LIWC-2007-based distance metric between two arbitrary sets of documents. The algorithm works as follows:

1. For each set of documents, combine all of the words into a single document
2. Calculate the percetage of words that reflect each of the 64 LIWC-2007 categories for each set's overall word counts
3. Take the Jensen-Shannon distance between the two 64-dimension vectors.

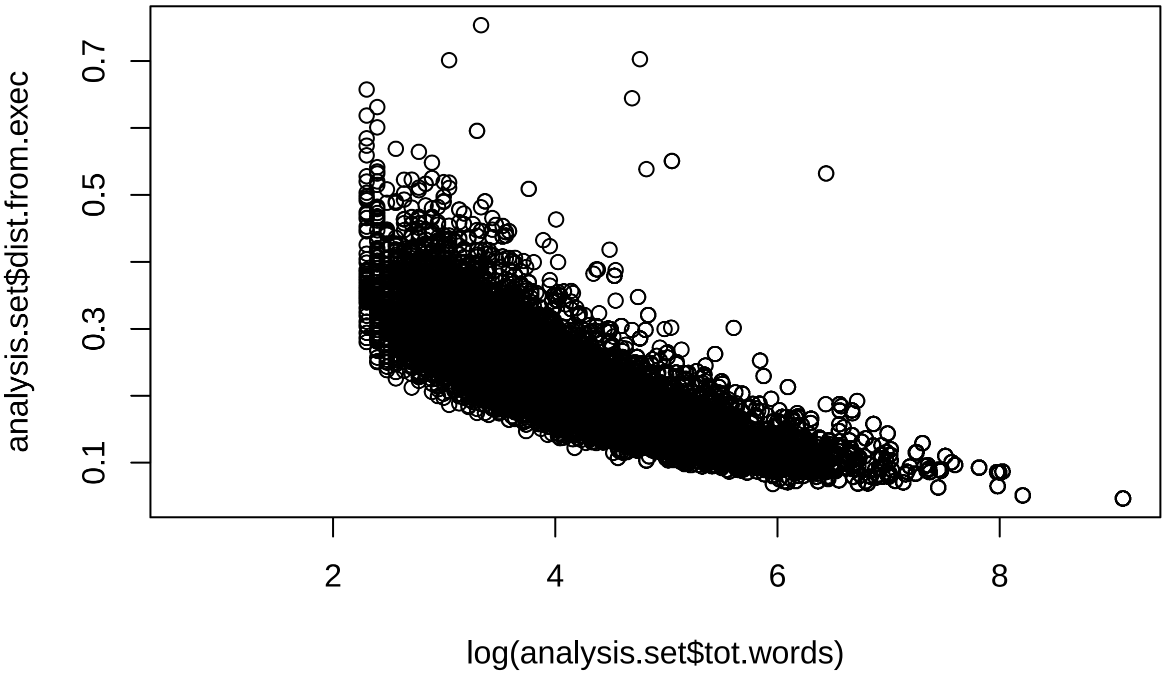
I applied this to the IBM data by agreegating all documents over 10 words long to the user level as well as creating a single document set from the 2898 comments written by the 878 executives who participated in the Worl Jam. This yielded 12648 users writing 28,683 comments over 10 words in length. I calculated the distance from each of the 12,648 users to the overall executive set and then incorporated these distance measures into the regressions we've been working with.

I could not find a configuration where the distance metric produced significant results.

Limitations compared to Goldberg et al.

The biggest problem here is that we have way less data than Amir did and thus end up with relatively uninformative indications of where individuals fit in terms of cultural style. They limited their analysis to individuals with at least 100 emails in a time period in order to make sure that there was enough data to actually pin down a person's location in LIWC space. The majority of our people have only one post, which gives very little information about what they are like, culturally speaking. The problem is that people who speak less will therefore be measured as arbitrarily further from execs than people who speak more. This is most apparent if you consider a single person who, for sake of argument, communicates exactly like the average exec. If you sample 100 words from that person, they will be measured as further from the average exec than if you sample 1000 words from them.

I plotted the log of the total number of words a user posted in comments on the x-axis by the distance from the average exec posting on the y-axis. Most of what this measure picks up is the variation in how much each individual participated.



Results

The measure is non-significantly related to whether a post is successful (p=0.38) across the models, showing a weakly negative relationship. This would be promising, save for the problems with uing the measure for users with low numbers of comments. I also limited the sample to just those users with 100 or more words, 250 or more words, and 500 or more words overall to see if, when looking at subsets with more data, the measure would be a reasonable indicator. None of these models produced significant results (fit without and with the topics).

Interestingly, log of the total number of words the user made in all posts, which is a necessary control to use this measure per the argument above, is also non-significant, a weak indication that posts from more prolific users were not more likely to be successful (see below for more on this, with regard to the number of forums in which a user posts).

Alternative analysis

A better approach here, since we can't reliably operationalize this at the individual level, is to focus at the aggregate level. Specifically, we want to do three things:

1. Find the distance between the quoted comments taken altogether and the exec comments taken altogether
2. Because we know that low numbers of total words strongly affects the distance score, we want to make sure we compare this distance to similarly-sized sets of comments. To do so, we generate 1000 bootstrapped samples with the same number of comments as the quoted set (274). For each of these 1000 sets of 274 comments, we compute the set's distirubiton across the LIWC categories and then calculate teh Jensen-Shannon distance to the set of comments created by executives.

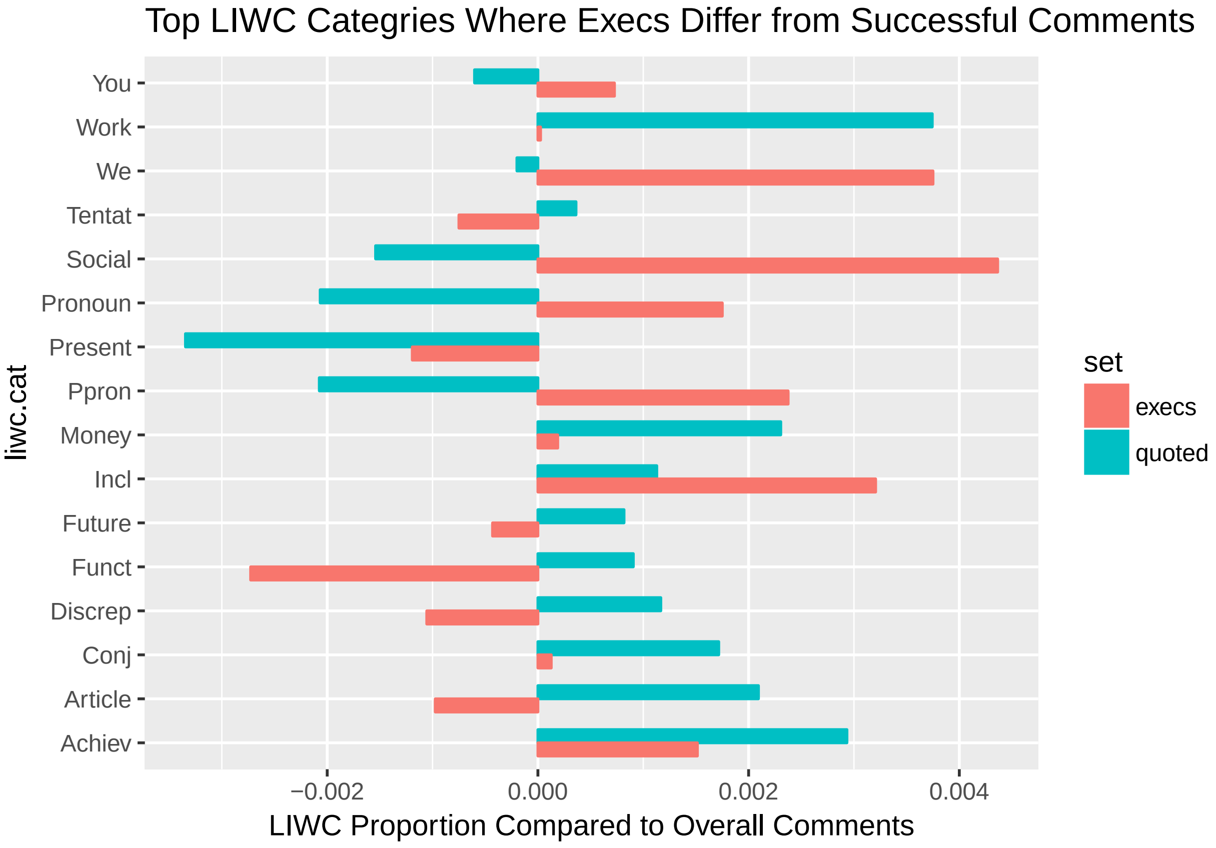
3. We use these bootstrapped distances to figure out whether the quoted comments are more or less like exec comments than random sets of documents.

Using the bootstrap approach I found, surprisingly, that quoted comments are less like exec comments in terms of LIWC scroe than a random set of other comments. It is not quite statistically significant (p=0.098), however. Cultural similarity, at least in terms of LIWC-categoriy similarity to the overall set of comments by executives, is not a strong indicator of successful comments.

It turns out that both the set of exec comments and the set of quoted comments are relatively far from each other as compared to random sets of 274 comments. I wanted to see if this was because only one or the other was further from random sets of comments. I pulled a random comparison set from the bootstrapped seta and used the other 999 sets' distance from it as a baseline. Both the exec set and the quoted set were the furtherest set of documents from the random comparison set.

So, execs have liwc distributions that are unlike random sets of comments, and quoted comments have liwc distributions that are different from random sets of comments, but both are also different from one another.

In order to compare how the exec and quoted sets differ from one another, I identified the the top LIWC-2007 categories on which their distrinbutions differ:


Top LIWC Categries Where Execs Differ from Successful Comments

Presentation Questions

1. <br> is a carriage return in html. It signals that the post was split into multiple lines. We can remove it for text display and replace it with an actual new line.

2. Structural Holes

- from multiple forums – I thought we had settled on structural holes being a longer-term theoretical concept? This would imply that we thought that participation in multiple forums was an indication that the person spanned different areas of the organization. This is perhaps a reasonable assumption, but we'd have to be explicit.
  - I ran a model with a dummy for whether a user posted to more than one forum. It has a slightly negative and completely insigificant relationship to whether a post is quoted. The zero-order relationship is also negative.
  - I am not sure what you mean on slide 26 by "same thing with multiple threads." Do you mean people who are posting in more than one thread? This gets tricky because it is will be highly correlated with how many times a person posts, which we might have to control for explicitly.
- from multiple locations in the firm – I'm not sure what you mean by locations? We determined before that we don't have good enough information on departmental location; figuring out which titles span mutliple departments would be even harder, so I don't think we can do much there. If you mean phyiscal location, we only have single locations for each individual. I did try a filter for people with "client," "customer," "consultiing," or "consultant" in their title to see if we could see any results from bridging the internal/external divide, but results of this dummy were non-significant in the overall model.
- Now that I see the comment on slide 26, I think you mean that there might be within-thread spanning of structural holes. This is an interesting idea, and strikes me as a novel hypothesis vis-a-vis Burt, that a conversation can span a structural hole rather than a person. I think this is (A) an exciting idea worth exploring and (B) will rely on the thread-level analyses to be doable. One of the things that's becoming clearer to me now that we have a theoretical framing is that past work has really emphasized individuals or projects as units of analysis, and not really been extended to consider "conversations" as sites where novelty can occur (unless you interpret the project in Uzzi or in Stark et al. as longer-term conversations form which successful ideas can emerge). If we can show this occuring in near-real time (not clear yet that we can), I think it'll be a large contribution.

3. Recombination

- I tested both hypotheses (with and without a squared term). Neither was significant.
- It looks like you'd prefer a quartile model for the recombination metric. Half of the posts have zero recombination novelty (no topic or a solo topic), so the comparisons will be between the 3rd and 4th quartiles and this lower half. I reran the model with a dummy for 3rd or 4th quartile of the novelty score, and neither coefficient was significant.

4. Cultural similarty

- see the first part of the memo
- we should discuss whether we are using the direct measure of some sort of qunatile transformation

5. Focus question

- That's correct: "focus.cos" is intercomment focus with the prior comment for those comments which are responding to another comment. "focus" is intracomment focus
- we should discuss whether we are using the direct measure of some sort of qunatile transformation

6. Multi-stage models

- in the prior memo, I presented two ways to do the multi-level models, which I thought unearthed an interesting story about how different topics can have different dynamics assocatied with them. This points, I think, to another hypthesis about how power is exercised through agenda setting.