Neumann, Linder, Desmarais:

Reading notes:

* Purposes: (1) How to extract plain text more easily and (2) Minimize the impact of boilerplate texts that tend to be quite different.
* Application – differences based on executive partisanship.
* I think the literature review in section 2 is interesting and useful.
* Table 1 and Table 2 aren’t well explained…nor is the link between the sentence on top of page 7 and the previous paragraph…
* “Finally, to remove words that are extremely rare (which also has the advantage of eliminating any remaining oddities) and thus add nothing substantive to our models while increasing their computational cost, we also discard any token types that occur in only one document.” – what does this mean?
* Why did you choose the threshold of ten for duplicates?
* What does the presence and frequency of words actually tell us?
* Refer to your tables, please!
* I like the final point in 5.1 about focusing the sample based on some sort of theoretical linkage.
* Law enforcement finding – timing of the data might matter- - BLM movement?
* Maybe describe the “fightin’ words” approach more clearly?

Comments:

I’m no expert in this topic, but think that it is interesting. As someone who has spent much time doing the “manual” approach, the idea of automating it seems appealing. However, I have some concerns:

1. What does this actually tell us about a local government? Is it more powerful than a careful case study-based analysis? Why? I know that case studies lack external validity, but at least a strong case could be made for partisan changes. Does your method improve the external validity so greatly that the internal validity becomes less of a concern?
2. Methodologically, it might make sense to eliminate words that do not show partisanship? Or to simply focus on a subset of potentially partisan words?
3. What does the lack of context mean? To me, it seems to be quite significant…
4. Given the other paper, how do you account for non-partisan local governments?

Thoughts from discussion at SPSA:

* Can you do changes over time rather than a cross-section? Upcoming election in 2018 provides chance to do this, most likely.
* Do you want to compare with non-partisan websites? Might also be interesting…
* I know that the innovation here has to do with boilerplate language, but does boilerplate actually tell us something of preferences? Perhaps is it worthy of some analysis?
* Can you control for city differences in some way? For example, the second model allows for covariates – perhaps include a few more than you do?

Other notes:

Tables 1 and 2 reference data on 'before', 'after' and 'current'. Before and after don't actually make much sense anymore, since they were in reference to the wayback machine data we're not using any more.

Figure 1 -- the word topic probabilities. This is using the results from LDA, which we're currently not using. It would still be a useful measure of the degree to which the boilerplate removal helps, but only when compared to the same plot if the boilerplate removal is not done. I don't think it is particularly informative on its own.

Figure 3 doesn't exist (i.e. there is a caption but no plot). This is the plot on the stm I referred to in my last email - i.e. the one that was purely diagnostic and not particularly useful any more - so at some point, I removed the plot, unfortunately not the figure.

Tables 9 and 10 are also on LDA.

Tables 11 and 12 (descriptive statistics) are partially on LDA, and the rest also isn't up to date any more. The most recent tables with descriptive statistics would be govWebsites/paper/presentation/descriptiveStatsIN.tex and descriptiveStatsLA.tex