Gmail **Bruce Desmarais <bruce.desmarais@gmail.com>**

Re: Political Analysis - Decision on Manuscript ID PA-2018-124

1 message

Bruce Desmarais <bdesmarais@psu.edu> Wed, Dec 12, 2018 at 3:34 PM
To: Fridolin Linder <fridolin.linder@nyu.edu>
Cc: Bruce Desmarais <bdesmarais@psu.edu>, Markus Neumann <mneumann.polsci@gmail.com>

Thanks for following up, Frido. I could do any time between 9 and 11 on Monday.

Bruce A. Desmarais
DeGrandis-McCourtney Early Career Professor in Political Science
Director, Graduate Programs in Social Data Analytics
Associate Director, Center for Social Data Analytics
Pennsylvania State University
brucedesmarais.com

On Wed, Dec 12, 2018 at 11:00 AM Fridolin Linder <fridolin.linder@nyu.edu> wrote:

> Hi all,
> Sorry for my late reply. Markus and I met about the paper on Friday and started a doc to reply to reviewer comments (you should have gotten an invitation). Should we schedule a call to talk about further steps soon?
> I will be traveling to Germany tomorrow so next week would probably be best for me. Here are some times that would work for me (all EST):
>
> Mon 12/17: 8am - 11am, 1pm - 2pm
> Tue 12/18: 8am - 2pm
> Thu 12/19: 8am - 2pm
>
> Would any of these times work for you?
>
> Best,
> Frido
>
> On Fri, Nov 30, 2018 at 12:58 PM Bruce Desmarais <bdesmarais@psu.edu> wrote:
>
>> Not the best possible outcome. But also not the worst. I think we should consider the note version that Gill suggests, as any . The reviews are actually quite positive relative to typical review outcomes at PA (the overwhelming majority of which result in rejections).
>>
>> To revise this into a note I think we need to do four things in addition to cutting the paper way down.
>>
>> 1. Make revisions and write a memo in which we address any substantive criticisms raised by the reviewer.
>> 2. Situate our methods contribution as a processing method for taking a directory full of heterogeneous file types from different government websites and producing a plain text dataset suitable for off-the-shelf text analysis methods.
>> 3. Put some subset of our methods into an R package (maybe the boilerplate removal).
>> 4. Describe and indicate that we'll release the existing data that we've collected.
>>
>> Two things then...what do you two think of this path? Would you like to meet during the week of the 10th to discuss a plan for revisions?
>>
>> -Bruce
>>
>> ----
>> Bruce A. Desmarais
>> DeGrandis-McCourtney Early Career Professor in Political Science
>> Director, Graduate Programs in Social Data Analytics
>> Associate Director, Center for Social Data Analytics
>> Pennsylvania State University

> brucedesmarais.com
>
> On Fri, Nov 30, 2018 at 11:21 AM Neumann, Markus <mvn5218@psu.edu> wrote:
>
>> -------- Forwarded Message --------
>> **Subject:**Political Analysis - Decision on Manuscript ID PA-2018-124
>> **Date:**Fri, 30 Nov 2018 15:54:11 +0000
>> **From:**Political Analysis <onbehalfof@manuscriptcentral.com>
>> **Reply-To:**jgill@american.edu
>> **To:**mvn5218@psu.edu
>>
>> 30-Nov-2018
>>
>> Dear Mr. Neumann,
>>
>> I write you in regards to manuscript # PA-2018-124 entitled "Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States" which you submitted to Political Analysis.
>>
>> We have reviewed your manuscript, and I have read it as well. Based on my evaluation of your manuscript and the external reviews, at this point in time I must decline it for publication in Political Analysis and close the file on it for further consideration. While the reviewers are critical, they are not overly critical but identify some important issues. You will see this below. My view is that this not a sufficiently novel methodological contribution to merit publication in Political Analysis as is. Specifically, the manuscript is a very useful guide to webscraping in a very particular context, but it is more of a tutorial nature than a research contribution. I am also reluctant to publish particular software guidance in a permanent venue. Having said all of that, I would welcome a de novo contribution in the form of a letter (per our guidelines) with the specific software details furnished in an appendix to be placed online if published (or preferably as an R package automating some of the decisions). Of course this letter would have to be written such that there is a methodological contribution independent of the appendix.
>>
>> I realize that this is not the outcome that you desired, but I hope that you find the advice of the reviewers, and my comments, helpful as you continue your research in this area.
>>
>> Sincerely,
>> Jeff Gill
>> Editor-in-Chief, Political Analysis
>> jgill@american.edu
>>
>> Reviewer(s)' Comments to Author:
>>
>> Reviewer: 1
>>
>> Comments to the Author
>> This paper explains the design of their website scraper for municipal websites and the results of text analysis of their content, but its argument is not strong enough methodologically or substantively.
>>
>> If the authors want to treat the data collection pipeline as a research output, they should make it widely available in the form of R packages, although it would be just a collection of other packages. If this is a methodological paper, they should present systematic comparison of different strategies, methods and tools.
>> I agree that extraction of plain text (body text extraction) is almost always the problem in online data collection, but author resort to manual coding and supervised machine learning to do the task. Since manual coding is not always the option, the authors should have tried to develop semi-supervised or unsupervised method. Creation of such a tool is certanly a great contribution.
>>
>> I think the use of Selenium is appropriate because it replicates humans accessing website. However, I do not understand why the author used wget to download pages. There must be dynamic elements in the pages that are generated or inserted by javascript. I consider this is a deficiency of the pipeline, which leading to incomplete data collection. Also, there is no need to use Selenium only to handle redirections.
>>
>> The authors removed boilerplate expressions line-by-line after converting files into plain texts, but I am not sure how line-by-line elimination is effective, because inline elements (e.g. span tags) can occur in the same line as substantive content. They should have considered the nested structure of HTML documents for effective body text

extraction.

It is strange to remove tokens before applying POS-tagger (spaCy), because accurate lemmatization requires syntactical parsing of the original texts.

Reviewer: 2

Comments to the Author
The manuscript is written in a clear language and well organized, and as a result easy to follow. The manuscript presents (1) a toolchain (pipeline) for automated collection and processing of English language government websites; the toolchain is built from elements readily available as open source software; (2) a corpus extracted from US local government websites;
(3) an analysis of the content of the corpus and its relationship to partisanship.
The main contribution is in (1), and the manuscript specifically emphasizes boilerplate (non-informative text) removal.

I read the manuscript as a how-to guide with substantive illustrations from one subfield. However, the substantive component takes more space than the methodological one. I think the manuscript would be better suited for PA as a letter focused on the toolchain and explicitly aim at a broader audience than researchers working with local government documents. Revised into that form, it would work better if it would give more information on why and how the various components of the toolchain are preferable over the alternatives, and what are their limitations. Perhaps the most important instance are the various methods for boilerplate identification, currently relegated to footnote 11. Here, it would help the manuscript's case a good deal if it could show how boilerplate removal matters for substantive conclusions, and show this not only for the analyzed corpus but also for some published findings.

Should the manuscript be revised along these lines, the question is how much would it contain that isn't already in textbooks or freely available course materials. Just the same, how-to papers written in a form accessible to political scientists sometimes become widely read. I think the manuscript has the potential to become such a paper if revised accordingly.

Reviewer: 3

Comments to the Author
This article, "Government websites as data: A methodological pipeline with application to the websites of municipalities in the United States" outlines a technique for gathering and content analyzing the data for municipal websites in the US. This article is a useful set of practices to help create a data-source for local politics. However, the way that it's written right now it seems more appropriate for a note than a full article. The article is written more as a narrative of what the authors did in this one case, using this technique for municipal websites, rather than a discussion of how this could be used with other types of websites and/or why websites may be a good source of data in local politics. Gathering data in local politics is difficult and recent efforts such as the LEAP project (https://na01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.leap-elections.org%2F&amp;data=02%7C01%7Cmvn5218%40psu.edu%7C6a329743e0484d6cb32708d656dc1499%7C7cf48d453ddb4389a9c1c115526eb52e%7C0%7C0%7C636791900578853822&amp;sdata=j41kpHxMdDLi8Fbhdu1BIErYBBLzVZoxVxxUa%2Bb8Hnc%3D&amp;reserved=0) and Sumner, Farris and Holman's crowdsourcing method (https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdrive.google.com%2Ffile%2Fd%2F19U0IxZhieNmfpuFOqTp276M_PsgUCagJ%2Fview&amp;data=02%7C01%7Cmvn5218%40psu.edu%7C6a329743e0484d6cb32708d656dc1499%7C7cf48d453ddb4389a9c1c115526eb52e%7C0%7C0%7C636791900578853822&amp;sdata=JGq%2By3kOzlu0sGIbAonPA3ShBV4TBp3T%2FduIXMO9LVQ%3D&amp;reserved=0) aim to make that process more access to scholars studying local and state politics. This website method could supplement these efforts and help scholars in the US (and abroad) study processes at the level where most people interact with the state – at the local level. There also seem to be other applications of being able to scrape and categorize more information on websites for scholars who may be interested in things like economic development in an area, responses to disasters, etc. and it would be useful to hear more about those types of applications. The authors talk about the improvements over other methods that they are offering, but are there drawbacks as well? The topic modeling has no valence attached, so while the modeling can speak to broad categories that are being mentioned, can it also enable scholars to say something about credit claiming or blame attribution? Can we say something about the absence of terms across different types of municipalities? Right now, the article uses the partisanship of the mayor as a way to test whether websites across municipalities differ. This doesn't seem to be the most useful example since most offices that people vote for in the US are non-partisan and the data here can speak to a great deal of other things that may define information cities share with citizens and the broader public. If the authors want to stay with the local politics example, it would be helpful to have a running example of a city throughout the text to show each step along the way. In addition, the topic model technique should allow the authors to say something about

> > the topics that occur in websites along multiple dimensions that they already have measures of like city size, region, median income. To that end, Table 5 could be changed to show something like the top 5/10 topics that occur by the partisanship of the mayor and then additional graphs or tables can the top topics by size of city, etc. Can the authors show how much overlap there is on topics and the probability that a topic will occur in a website? It would also be helpful to have a table or a figure that lays out the steps from pre-processing to topic modeling along with all of the R packages so that it's more of a user guide for readers.
> >
> > Sumner, Jane Lawrence, Emily Farris, Mirya Holman. "Crowdsource Reliable Local Data" presented at the 2018 Political Methodology Meeting
>
> --
> Fridolin Linder
> Postdoctoral Associate
> Social Media and Political Participation Laboratory
> New York University
> fridolin-linder.com