Notes of the user group meeting on Jun 13, 2014

Notes from the June 13, 2014 HTRC User Community teleconference call

Present:

Rachel Brekhus, University of Missouri

Matt Wilkens, University of Notre Dame

Michelle Paolillo, Cornell University

Ted Underwood, Loretta Auvil, Sayan Bhattacharyya, Peter Organisciak, University of Illinois at Urbana-Champaign

Miao Chen, Indiana University

(This conference call focused on the recently released alpha version of "feature extraction" from the HTRC. For specific details of that release, please see **this announcement**, and to get a sense of the broader context for the release, please see **this blog post**.)

Discussion summary notes:

From Matt Wilkens: *“As a user, one of the first things I think I will be looking to do, would be to pull all the feature counts into a database, Has there been any thought on the part of HTRC about putting the feature counts directly into a database instead of into a file?”*

Discussion: Ultimately, HTRC is thinking of providing (some type of) API(s). So, instead of downloading files, access would be provided with API(s') access.

Peter asked: “How much interest may there be in engineering a more complex API? What other things can we (prioritize to) do with the data? Would the likely eventual size of the data (multiple gigabytes of data per decade), when we scale up to the contents of the entire HTRC corpus, end up exceeding what most regular people can download to,or work with, on their desktop machines? Of course, people may want to do things large-scale. Would it be reasonable or unreasonable to push that effort on to the user?”

Discussion: You can actually rsync on the individual volumes as well as the grouped tar files. However, the volume IDs have to be filename friendly, which HT volume IDs are not. We could provide a mapping from the HT volume IDs to filename-friendly IDs. There are reserved characters used in HT/HTRC volume IDs — characters that should not be used as part of names in the file system. A volumeID is not equivalent to a filename.

Matt commented that “rsync was great!”

“Can we do prepared filename lists, for predictable kinds of worksets? Can we organize/ split-up datasets by genre? An argument for doing so is that, for example, fiction may be a high-demand-for-download genre compared to other genres.”

Discussion: We debated about how to group the datasets, and finally we decided to use year/decade/chronological type of information for grouping the datasets — mainly because this would be less controversial. For example, questions like “What counts as fiction?” could be highly controversial — the present grouping by chronology lets us bypass that controversy. However, this is not wholly unproblematic either — as (e.g.) a book listed as having been published in 1916 could well be a reprint of something published in 1872, etc.

“This is going to be a very new kind of thing for many users. How to describe to users how to use this tool? What kind of documentation is planned?” Rachel asks for examples and documentation.

Discussion: We need to get some articles out there (showing how the extracted features can be used), and also create some tutorials for users. For people doing this kind of work for the first time, it will also be useful to have an introduction to: ‘How to use this data with Python’, ‘How to use this data with R’, etc. However, for those users who may not know how to program, to go straight to slicing and dicing such large datasets may be too big a step. We could try to provide initial maps of page structures for at least the English-language volumes. Since this would involve only metadata, we should be able to do this for both Google-digitized and non-Google digitized volumes. It would also be useful to have a tutorial to show people how to interact with the JSON, how to use rsync, etc. Is providing examples being planned? For someone who uses Excel in his/her work, it will probably be more useful to have examples along the lines of ‘How can this task X be performed using Excel?’, rather than providing an “Excel manual.”

“One of the features currently being provided is: beginning-of-line and end-of-line characters.” Matt asked, “What is the use case for this?” Ted responded, indicating that it provides the ability to classify pages, such as index, table of contents, identify poetry, etc.

Discussion: Lines on indices, as well lines of poetry in 19th century literature, etc., often begin with capital letters. Lines in an index often end with numbers (numerals). So, this information is useful for page-level classifications. However, it does take a lot of disk space to store these beginning-of-line and end-of-line characters! Luckily, things will eventually be set up in such a way that, if you are interested in only a part of the features, you will not need to download all of the features — that is, if you don't need beginning-of-line and end-of-line characters, you will eventually have a way so that you don't have to download them when downloading other features. We can also potentially develop ultra-slim, slim and deluxe, etc., versions of downloadable features. In addition, if features get to be provided on a volume-level rather than on a page-level basis, then the disk space needed will be significantly smaller.

"Should all features be in one place, or should they be combined from different places? Going back to one of the issues underlying the question of databases versus files (that we discussed earlier in the meeting): storing multiple files per volume takes up quite a bit of disk space. If you are interested only in part of a feature, should you still have to download (eventually) all the features nonetheless?"

Discussion: Dimensionality reduction could mean reduce all to classes of begin/end of line characters, to number of uppercase, lowercase, number, punctuation (all included), and not maintain the details separately. We could also reduce to token counts for the volume and perhaps exclude the running headers. We did hear that this could be very useful.

Using the extracted features twill require some effort on the part of the user. The number of users who have been waiting to try to do this is likely to be small. The bottleneck is not so much the HTRC, but rather the smallness of the number of users.

Discussion: When people will see what they can do with the features, and when they see that it is tractable, then there will be more acceptance for this.

From Rachel: There's a tool ("Paper Machine") that came out recently from the American Sociological Association [?] They are doing some extraction, from government documents about policy statements. They are trying to do data visualization, some kind of word cloud out of it.

Discussion: Are government documents in the HT corpus always clearly marked as such? Not sure.

Would it be useful to do entity extraction (named entity recognition, NER) with these features?

Discussion: Experience shows that doing NER with personal names is relatively straightforward and does better than with place names, which only have a 70% accuracy (improvable to about 80% with nation-specific stuff and hand-massaging). By contrast, POS tagging is well over 98% accurate. It is surprisingly hard to get multiple human beings to agree on what a "place" is or is not. There is also a "public relations" angle to be considered here. If you do so, you would have put a data set with a lot of errors out there, or that is how it will be perceived (even if we put out the precision and recall numbers, or include a warning that NER is error-prone).

Matt uses Google for identifying/validating locations. which generates quantitatively better results than the GeoNames data. However, while it is fantastic to have NER, it would have to be reviewed by hand to improve accuracy. Perhaps this could be a Version 2 feature. Yet another possibility could be to use multiple NER packages, and report only those named entities for which there is agreement from multiple NER packages.