Briefing Report for WE1S Advisory Board

(document created 20 July 2019; last rev. 25 July 2019)

This briefing report for the WE1S Advisory Board prepares for the WE1S project's meeting with its Board on Friday, 2 August 2019, 11am-1pm, Pacific time (via Zoom virtual conferencing at meeting location https://ucsb.zoom.us/j/760-021-1662). In the meeting's first hour, our summer research camp teams will give lightning presentations of example research findings. In the second hour, which will be a plenary consultation, WE1S would like to ask the Board's advice about some or all of the strategy questions posted at the end of this report.

—Alan Liu, Jeremy Douglass, Scott Kleinman, Lindsay Thomas

***Note**: Links to some resources mentioned below require a login, which we are sending separately by email.*

Highlights of recent project work

Corpus collection

On the recommendation of its Advisory Board (at last summer's meeting), WE1S concentrated in academic year 2018-2019 on collecting documents from U.S.-only media, though we are now also finishing gathering from top-circulation British and Canadian English-language sources for possible future analysis. At this time, our corpus consists of about 2.3 million unique documents, including documents from news and magazine sources, government institutions and foundations, Reddit, and Twitter, dating mostly from 2000 to the present (with some documents going back to 1981 with the onset of fully digital articles from *The New York Times*). These materials came from about 880 different sources, including:

- mainstream, regional, local, and student newspapers and magazines (mainly from the LexisNexis and ProQuest databases)
- born-digital online news sources (scraped using our "Chomp" tool)
- specialized online sources such as government and foundation humanities funding agencies and Humanities Councils
- and social media (Reddit and Twitter).

Full texts in our corpus (with those under copyright or from databases with contractual constraints later to be stored in "bagified" form only) have been assembled into different WE1S "collections" shaped to facilitate asking different kinds of research questions. At present, the core WE1S collection (*"us-humanities-all-no-reddit"*) consists of about 83,000 articles and other U.S. texts mentioning "humanities." (Reddit posts, which we are finding to be very telling, are excluded at present from the core collection because their scale tends to overwhelm our ability to pre-process, model, and analyze. Our Twitter scraping is recent, so we are also studying it separately.)

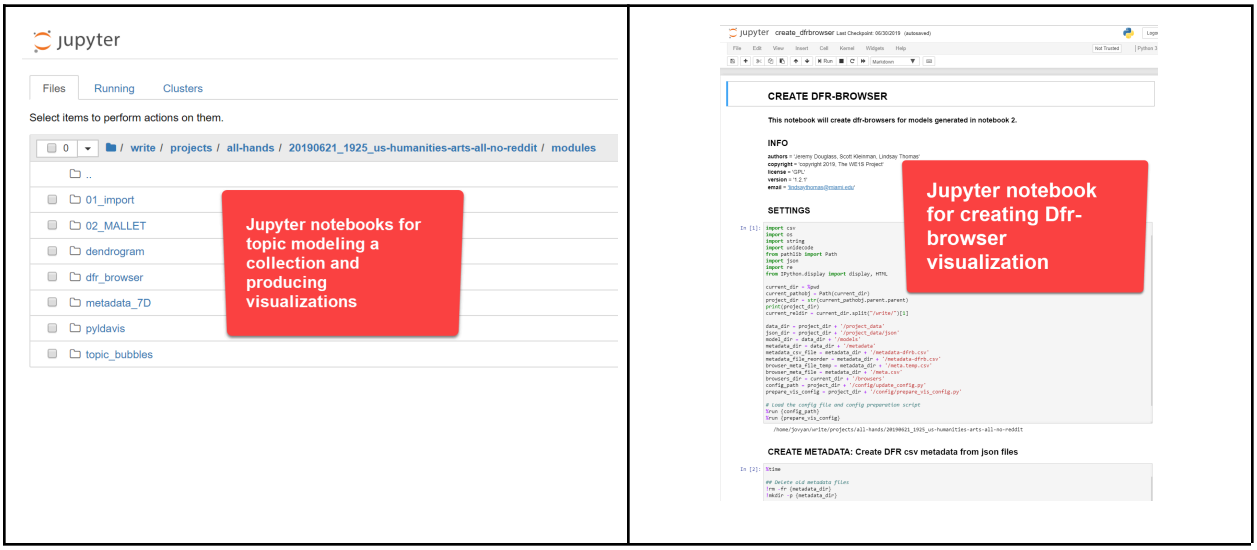
In addition, we have collections of texts mentioning “liberal arts,” “the arts,” “science(s),” and “literature” as well as a down-sampled “random” comparison corpus (collected by searching on common English words). We mix-and-match all these collections as needed for particular research purposes. *(See WE1S Collection Census for more detail.)*

Our collections have been pre-processed through steps that include de-duplicating articles, regularizing metadata, and stopwording (as well as some limited consolidation of named-entity and compound phrases). We have also enhanced our collections’ metadata (citation information) by hand-tagging sources for their geographical region; genre; media; and self-identified association with political, religious, social-group, and educational sectors. *(See WE!S Source Metadata Tags.)*

In addition, our Interpretation Lab has “wikified” one version of our core collection by using the Illinois Wikifier (for context-sensitive disambiguation of named entities via Wikipedia-article matching). One goal is to try to improve the coherence of topic models. Another, original to our project, is additionally to harvest through wikification geolocation and date information for named entities mentioned in articles. This has allowed us to create an experimentally enhanced version of our standard topic model that we can analyze with new tools. *(See below under “Models” and “Topic Model Observatory.”)*

Technical platform

We advanced our cloud-based, containerized WE1S “Workspace” for reproducibly collecting, pre-processing, topic modeling, and visualizing our collections. At the heart of the Workspace is a series of Jupyter notebooks that modularly make and visualize topic models. We also advanced our WE1S “Manager” for managing, querying, and working with our collections through a schema-based “manifest” system operating on a MongoDB database of our collections (though this has not been finalized).



Modeling

We topic modeled our core collection (*us-humanities-all-no-reddit*) at the granularities of 25, 50, 100, 150, 200, and 250 topics. *(See model and visualizations.)* After optimization testing (via Generalized Latent Semantic Analysis applied to LDA), we focused on the 250-topic model as our default.

We have also made about twenty other models based on various combinations of our collected materials--e.g., models of collections made by searching on both "humanities" and "science," models assembled from ProQuest's Ethnic NewsWatch or GenderWatch, and a model based on online materials from government agencies, non-profit organizations, state Humanities Councils, and other organizations funding or supporting the humanities in the U.S. *(See WE1S Topic Model Registry.)*

In addition, as described above, we created a special, experimental version of our default topic model that has been "wikified" and whose named entities (when they can be matched to Wikipedia articles on those entities) are geocoded and dated. The GeoD and TimeD visualization tools we created for exploring this model allow us to ask research questions about the "geographical profile" and "historical profile" of topics--i.e., what locations or historical dates a topic is associated with. *(See experimental model and its visualizations.)*

"Topic Model Observatory" visualizations

We implemented a suite of interactive visualization tools for exploring topic models that we call our "Topic Model Observatory."

These include adapted versions of Dfr-browser and pyLDAvis, which were created by others (Andrew Goldsworthy on the WE1S Advisory Board created Dfr-browser). (We also use the pyLDAvis interface for alternative purposes, such as seeing what publication sources are associated with topics through our Metadata-pyLDAvis tool.).

New visualization tools we have created include TopicBubbles (especially intuitive for looking at multiple topics together), DendrogramViewer (for inspecting possible clusters of topics), Metadata7D (for looking at the association of topics with sources tagged by custom metadata), GeoD (for seeing on a map the "geographical profile" of topics based on geocoded named entities), and TimeD (for seeing the "historical profile" of topics based on dates mentioned in documents). *(See our Topic Model Observatory Guide for FAQs, screenshots, and working demos of most of these tools. See also the visualizations for our experimental wikified model.)*

Topic Model Interpretation Protocol

We created (and are currently using version 2.0) of our topic model Interpretation Protocol, which standardizes and documents steps for model interpretation. The Protocol consists of modules of instructions, observation waypoints (features of a model to look at, tools to use), and analysis procedures. These modules are coupled together in flexible workflows for particular

research questions. Typically, an interpretation workflow (preceded by free-form exploration and note-taking activities that we are now gathering “best practices” for) consists in taking an initial overview of a topic model, looking closely at keywords and topics in a model, comparing keywords and topics, and conducting analysis/synthesis and reporting steps. These modules are implemented as Qualtrics surveys that deposit a transcript of interpretation activity as part of the dataset supporting interpretation results. (The Qualtrics surveys are usable by others as exported .QSF files or, for those without an institutional Qualtrics license, HTML forms [pending]. They are also available as reference copies in Word and PDF format that can be filled out as ordinary documents.)

Note: the WE1S Interpretation Protocol does not presume that there is a one-size-fits-all interpretation process for topic models. Instead, the general goal is to show that digital-humanities research requires *some* regularized, documented, reproducible, and extensible workflow for humans learning from machine learning. When eventually released in a GitHub repo, the WE1S topic model interpretation protocol can be forked, modified, and extended by others for other projects. We hope that the digital humanities field will evolve families of related protocols for interpreting different kinds of machine learning. (We are also tantalized by the possibility of extending the interpretation-protocol approach to “traditional” humanities interpretation--e.g., through a set of modules for what to do and look for [and how to document the process] while “close reading.”)

Research questions addressed to date

Much of WE1S research work in its first 1.75 years has been preparatory: developing a technical platform; creating collections; and researching contexts, sources, materials, and issues related to journalism, the media more broadly, and the humanities. (For example, see our “scoping” and “area of focus” reports.) Because we did not start with a pre-made corpus, just collecting materials at scale--and making the technical platform and intellectual context for doing that well--has been challenging.

However, as of summer 2019 we made the pivot from development work to answering research questions preparing for our grant deliverables: analyses of public discourse on the humanities and suggestions for humanities advocacy. Our 2019 summer research camp, which includes over 30 faculty and graduate-student participants from a combination of UCSB, CSUN, and U. Miami (with additional participants from UCLA, Texas A & M, Claremont Graduate University, and Illinois Institute of Technology), is thus working on what we think are important research questions that can be addressed using our collections, topic models, tools, and interpretation protocol. These research questions, which originally arose through iterative brainstorming in earlier stages of the project, are organized in the following “research bands” (like radio signal bands grouping together channels of inquiry):

- Public discussion of the humanities “crisis.”
- Student discussion of the humanities.

- The geographical map of who and what is mentioned (and from what sources) in public discussion of the humanities.
- How underrepresented social groups figure in relation to the humanities in public discussion.
- How the humanities are related to issues of economic value in public discussion.
- The broader social profile of the humanities by comparison with science in public discussion.
- How different media forms (print, online, social media, TV and radio [in transcript form]) inflect public discourse on the humanities.
- The impact of funding agencies (e.g., NEH, Mellon) and humanities organizations (e.g., state Humanities Councils) on public discussion of the humanities.

We anticipate that much of academic year 2019-2020 will be devoted to additional cycles of work addressing high-value research questions in these and other "bands."

(The summer research camp also has teams at work on planning for WE1S's final advocatory and analytic outputs, project documentation and dissemination, and continued technical work on modeling and visualization.)

Human Subjects Research

We sought and received approval from the Mellon Foundation to add some limited "human subjects" research to our originally-defined research methods. To evaluate how the discourse on the humanities we have collected from print, online, and other media sources matches (or contrasts with) that of actual students and others in our own community responding directly, we conducted at UCSB (with IRB approval) an online survey of 40 questions about the humanities and higher education. Among undergraduate students, 125 (including 46 who identified as first-generation to college) started the survey and 97 completed it. Among non-undergraduates in the UCSB community (graduate students, faculty, and staff), 122 started the survey and 76 completed it. In addition, we conducted three focus-group meetings with students and others. Results from these research activities are currently being analyzed as part of our summer research camp work.

We note that our human subjects research is not Intended to be wide-scale or scientifically sampled. Its purpose is to assist us in designing and interpreting our main efforts in data collection and analysis. We are looking for ways that direct response from students and others can surface features in our topic models that we might otherwise miss (e.g., keywords to analyze). Also, by grounding our big-data approach in close reading of a campus's direct responses about the humanities, we hope we can particularize the "public" in public humanities and plan for what humanities advocacy might look like in local communities.

We also note that our project's Curriculum Lab has complemented our human subjects research on student views by innovating (and documenting in research blog posts) experimental courses

at UCSB. These courses ask students to think across the divide between the humanities and sciences, and between the academy and other social contexts.

Questions our Advisory Board can help us think through

The above report is background for WE1S's lightning presentations to our Advisory Board on August 2nd (11am - noon, Pacific time) and our plenary consultation with the Board afterwards (noon - 1pm). (Reminder: we are meeting with our Board via Zoom virtual conferencing at meeting location https://ucsb.zoom.us/j/760-021-1662)

We close this briefing report by hanging in the air a few questions about strategy going forward that we could use help with. (It may not be possible to get to all of these in our consultation session on August 2. But simply outlining the questions can help structure our plenary consultation session):

1. We are finding it a challenge to find good ways to identify and analyze in our collections/models (including those based on ProQuest Ethnic NewsWatch and GenderWatch sources) how underrepresented and first-generation groups are discussed in the media in relation to the humanities. **How can we enhance our research collections or methods to address this better, or to ascertain whether the conversation about such groups *in relation specifically to the humanities* is actually occurring in meaningful ways in public media at all?**

2. As described above, one experimental version of our core topic model has been "wikified" so that named entities are disambiguated through automated reference to Wikipedia articles and also associated with geolocation and chronological data. We are aware, of course, of well-known criticisms of Wikipedia for failing to provide "an equal, wide survey" of people and the world (to borrow the subtitle of one of John Barrell's books: *English Literature in History, 1730-80: An Equal, Wide Survey*). More generally, we are also aware that our underlying collection(s) of media materials is not "an equal, wide survey." **What is the best way for WE1S to address the issues of omission, underrepresentation, and unbalance in studying "public" discourse when we release findings next year to the public and also to the scholarly community?**

3. **How might WE1S integrate its research into public discourse on the humanities with statistical research into the state of the humanities of the sort gathered by Humanities Indicators?**

4. **What are the best kinds of advocacy outputs that WE1S, using its research, can suggest (or prototype) for advocating the humanities in education and society?**