DAHC Proposal

CFP: https://goo.gl/bbNnQQ
This Proposal Template: https://goo.gl/xdkWUe
Submit: dahc@hfa.ucsb.edu

"WhatEvery1Says" Text Analysis Hackerspace

A Proposal for the DAHC

***Lead Faculty Member(s)*: Alan Liu**
***Application Date*: Dec. 10, 2017** (document last rev. 12/10/17)

Contents

(A) Project Summary:

Description of project and space use concept. This should address goals such as research, teaching, and/or reaching broader publics. How will the project benefit from being in the DAHC, as a community and as a physical space? You may include URLs and digital resources, including current or past work or portfolio items relevant to the proposal.

Background - Part 1

WhatEvery1Says (WE1S) began at UCSB in 2013-2017 as a prototype project focused on using text-analysis methods to study public discourse about the humanities (especially in journalism) at big data scales (prototype development site). It gathered about 36,000 articles from major U.S. newspapers and other journalistic sources dating from 1981 to 2014, and analyzed them using the machine-learning method of "topic modeling" (increasingly used in the digital humanities and digital social sciences [definition below]).

Recently The Andrew W. Mellon Foundation granted the WE1S project $1.1 million for three years (Oct. 2017 to Sept. 2020) to expand the scope of its research and technical methods (UCSB announcement). The grant will allow WE1S to scale up its collection of public discourse about the humanities to a more representative corpus across regions, nations, states, and cities; and to evolve its combined computational and human-interpretation methods for analyzing how the public understands the humanities. Another core mission will be to study how racial, ethnic, gender, first-generation student and other groups are positioned by the media, or position themselves in the media, in relation to the humanities. WE1S's end goal is to produce analyses and materials for different audiences--the media, politicians, parents, students, and scholars--that can lead to a more robust, diverse, and informed discussion of the role of the humanities in public life. (For more information, see WE1S Prospectus.)

Background - Part 2

Meanwhile, during the same 2013-2017 period when WE1S was incubating, a critical mass of graduate students and faculty with interests in advanced text analysis methods in the digital humanities, computational linguistics, and natural language processing has been developing at

UCSB. These researchers (those currently known are named below under "Participants") are scattered among the English, Linguistics, Media Arts & Technology, and Sociology programs. Members of this group have met informally and occasionally, including at meetings or events of the WE1S project. A current shared interest of the group is "word embedding" (or "word vector") analysis, a state-of-the-art "shallow neural network" artificial intelligence approach to understanding the semantic relations between words in any language that holds the tantalizing promise of showing how a society semantically "thinks," complete with the cultural and other biases of that thinking. Up to the present, however, there has been no venue or organizational framework at UCSB that can help these text-analysis researchers collaborate and also disseminate their methods to others, including to both the campus research community at large and to the diverse group of women and men graduate students participating as research assistants in other WE1S tasks who are eager to be introduced to advanced text analysis methods.

Proposal

WE1S proposes to create a "Text Analysis Hackerspace" in the DAHC that can serve as the hub for faculty and graduate students with an interest in text analysis methods. Using its grant resources, WE1S will sponsor the participation in the hackerspace of an interdisciplinary group at UCSB with advanced text-analysis and linguistics skills. The group will consist of WE1S-funded graduate-student research assistants and Linguistics faculty member Fermín Moscoso del Prado Martín (with WE1S PI Alan Liu participating). Their role will be to incorporate in the main WE1S workflow a set of experiments in leading-edge text analysis methods; to introduce text analysis at both basic and advanced levels to other students and faculty; and to cross-fertilize with other DAHC and Wireframe projects (e.g., in regard to visualization methods for text analysis or the analysis of social media text). The goal is not just to aid the WE1S project itself (by advancing its research workflow and methods, and by training other research assistants less familiar with text analysis) but to contribute generally to research and teaching at UCSB. After all, so-called "distant reading" computational approaches to text are increasingly important across a range of humanistic and social-science disciplines--to the point that basic literacy in these methods can be an entrée to getting interviews for some jobs (as witnessed in the recent job-market experience of graduate students). Giving a text-analysis group a place in the DAHC would for the first time provide text-analysis enthusiasts at UCSB with a crystallizing physical site, provide them with an ideal demonstration and training space, and allow them to interact with other projects and methods incubated in the DAHC.

Specifically, the following is what the Text-Analysis Hackerspace would seek to achieve:

1. *Experiment with advanced text-analysis methods such as topic modeling and word embedding that are currently at the front line of digital humanities research (and incorporate these methods in a beginner-friendly "data notebook" workflow).*

   A leading method of machine-learning analysis, "topic modeling" discovers through statistical means the existence, relative weight, and distribution of “topics” across documents (where topics are represented as a probability model of correlated words often indicative of what a human see as "themes"). Widespread adoption and discussion of the method in the digital humanities and such other fields as the digital social sciences have demonstrated its usefulness. The current frontier of topic modeling research includes optimization of models, cluster and visualization analysis, longitudinal time studies, and multilingual topic models.

"Word embedding" is a so-called "shallow neural network" computational approach that models the semantic relations between words in a text corpus, thus making it possible to model not just the co-occurrence of words but logical relations of analogy and opposition between words. In an oft-cited example, word embedding mathematically derives such analogies as the following in a language: "king" is to "man" as "queen" is to "woman." In essence, word embedding reveals the relations of conceptual similarity and difference in language that illuminate how a society "thinks" about things, cultural biases and all. Word embedding lies at the current leading edge of text analysis, with further directions of research (e.g., longitudinal applications of such approaches) just beginning to emerge.

"Data notebooks" like Juypter notebooks (formerly iPython notebooks) allow live code to be run within what appear to be document pages that include explanations, instructions, and narrative. They are simultaneously a learning, collaboration, and operational tool. They are thus an excellent means for a group of collaborators at different levels of technical skill to work together on data-science projects. The WE1S project is creating a chained series of Jupyter notebooks that guide participants through the topic modeling process. The Text-Analysis Hackerspace would advance these methods by also extending the workflow to incorporate experiments in other kinds of text analysis such as word embedding.

The Text Analysis Hackerspace would function as a hub of collaborative experimental work in text analysis.

2. *Use the DAHC space to introduce other students and faculty at UCSB to text analysis.* An increasing number of faculty and students from across the disciplines are now interested in learning about text analysis--spanning across the range of basic digital research methods (e.g., use of Markdown and GitHub), established text-analysis methods (e.g., collocation analysis), mainstream advanced approaches (e.g., topic modeling), and bleeding-edge methods (e.g., word embedding). WE1S is planning workshops (open to all UCSB faculty and students) on the following:

   * Markdown & GitHub: Modern Digital Practices for Scholarly Research (originally scheduled for Dec. 12, 2017; to be rescheduled due to the Thomas Fire emergency in Jan. 2018) (See workshop page)

   * Beginner's Text Analysis

   * Topic Modeling

   * Word Embedding

   * Python and R for Text Analysis

A DAHC venue would allow WE1S to open its workshops to the broader campus community. A good precedent is the topic modeling workshop that WE1S ran on August 28, 2015, which attracted a surprisingly large group of both faculty and graduate students from multiple departments (see workshop page).

In addition, the Text Analysis Hackerspace would be used as one of the venues for WE1S's "summer research camps," which each summer during 2018-2020 will convene a group of approximately 12 research assistants and faculty to conduct intensive training, production, and interpretation work based on text analysis. (WE1S also plans for the summer research camps in 2019 and 2020 to include participant RAs from other universities.)

The Text Analysis Hackerspace would function as a demonstration and training space.

3. *Collaborate and cross-fertilize with other projects in the DAHC.* It makes sense for the WE1S Text Analysis Hackerspace to be in the DAHC because it would be able to cross-fertilize with digital visualization, social media, and other projects that bear integrally on its research. Just as one instance, much text analysis research relies on generic visualizations created in such programming languages as R with standard graphics libraries. But as indicated by the WE1S project's adaptation of D3.js in the dfr-browser exploration-interface for topic models, graphing is a bare minimum of what is possible in visualizing the results of text analysis. WE1S can collaborate with other projects (e.g., Wireframe) that are familiar with alternative, advanced ways of modeling and visualizing cultural knowledge.

   The Text Analysis Hackerspace will benefit from collaboration with other DHAC projects.

4. *Contribute an additional idea of "space."* The DAHC CFP calls for proposals that make "creative use of space." Of course, the Text Analysis Hackerspace will be a physical space in the DAHC (floorspace with workstation, etc.). But, beyond that, it will catalyze new intellectual understandings of "space." Some of the most advanced work on digital text-analysis at present--including especially "word embedding"--occurs in so-called "high dimension" conceptual spaces in which any word in a textual corpus has *N* dimensions reporting on its relations with all other words in the corpus (e.g., how often or to what degree it collocates with another word). Beyond the practical outcomes of such text analysis, the *theoretical* outcome is a robust rethinking of what significant (meaningful) "space" means. WE1S would like to collaborate with Wireframe and other proposed DAHC projects (e.g., the proposal from Prof. Marcos Novak, one of the early innovators in digital imaginings of "transvergent," 4-plus dimensional "space") in imagining and visualizing what high-dimension "spaces" needed for word vector, artificial intelligence, and other recent text analysis methods can be. In the process, it will contribute fresh insights to the whole discussion of what a collaboration and maker "space" like the DAHC can be.

   Much like the Spatial Studies Center at UCSB (or, in an earlier era, Project Alexandria at UCSB), the Text Analysis Hackerspace would join together the notions of mathematically-modeled abstract space and physical space in the intellectual collaborative maker space of the DAHC.

5. *Make/build a "deep learning" workstation for text-analysis."* Besides participating in the common spirit of the DAHC by contributing to the idea of "space," the Text Analysis Hackerspace group will also share in the "maker" or "builder" spirit of the DAHC by building a DIY workstation optimized for advanced text analysis, including for neural-network artificial intelligence machine learning. The making of this workstation will be documented through photos, videos, and write-ups for the WE1S blog. *(See under "Equipment" below for details.)*

(B) Participants

List any participating faculty / staff / students or others, along with their role in the project.

Faculty

- Alan Liu -- Professor, English Dept.; PI of WE1S Project (specializing in digital humanities)
- Fermín Moscoso del Prado Martín -- Asst. Professor, Linguistics Dept. (specializing in Psycholinguistics, quantitative linguistics, computational linguistics, cognitive modelling, statistics)
- [WE1S will also be hiring two postdoctoral scholars for academic years 2018-19 and 2019-20. It is anticipated that these postdocs will be participants in the Text Analysis Hackerspace]

Graduate-student Research Assistants (funded by WE1S)

The following graduate students have committed to being WE1S-funded research assistants affiliated with the Text Analysis Hackerspace. They will work self-regulated, flexible hours paced by team meetings, team events, and team workshops conducted for the larger research community at UCSB:

- Sandra Auderset (Ph.D. student, Linguistics)
- Devin Cornell (Ph.D. student, Sociology)
- Nicholas Lester (Ph.D. student, Linguistics)
- Fabian Offert (Ph.D. student, Media Arts & Technology)
- Teddy Roland (English; function as a member of Text-Analysis Hackerspace and also as DAHC GSR)
- Chloe Willis (Ph.D. student, Linguistics)

(C) Outline of Project Requirements

Floor Space

The DAHC floor plan is divided into units of 125, 250, and 500 sq-ft. Please indicate if you have a preferred area and/or location, and give a brief description of how you plan to use it. See CFP map.

The Text Analysis Hackerspace requests a 125 sq-ft space. For the purpose of the workshops and demonstrations it will run, it will utilize the common area in the middle of the DAHC.

Equipment

List any equipment (hardware or software) that you plan to install or maintain in the DAHC Space. Along with this, please indicate administrative or infrastructural support that this equipment requires, such as outlets, Ethernet ports/Wi-Fi, etc. Note that the DAHC is able to offer a modest grant for acquiring new equipment. Also, the DAHC will make available laptops with Windows and Mac operating systems that can be checked out by DAHC participants on request. Software can be installed on these machines, however memory will be periodically wiped.

The Text Analysis Hackerspace requests $1,000 from DAHC (to be combined with WE1S Mellon funding; see below) to create a "deep learning" (i.e., powerful machine-learning and artificial-intelligence-learning) workstation for advanced text-analysis work. This workstation will be created by its participants as part of a "maker" project in keeping with the builder/maker spirit of the DAHC.

One graduate-student participant in the Text Analysis Hackerspace has already prototyped such a deep-learning machine of his own based on the NVIDIA company's best consumer-grade GPU (graphics-processing-unit used for today's leading artificial-intelligence computing) supplemented by high-end SSD storage (fast solid-state hard drives). The parallel-computing nature of such a machine can accelerate several kinds of text-analysis work, including word embedding and NLP (natural language processing).

This "deep learning" workstation will be integrated/networked in an overall WE1S technical platform that includes physical campus workstations in the English Dept., virtual workspaces served from the English Dept., and cloud-based storage and processing services. Essentially, the "deep learning" workstation located in the DAHC will be the thoroughbred horse in the race: powerful, dedicated to the most cutting-edge work, but part of a stable of workhorses providing a robust technology platform.

The total cost of the "deep learning" machine is estimated to be around $2,000. WE1S has minor funding in its Mellon budget for hardware, which it will pool with a $1,000 bequest from the DAHC for equipment to create such a machine.

Access

The DAHC is open during Music Library hours (http://dahc.ucsb.edu/about), and all lead faculty members will be assigned a key to the Commons. If you anticipate further access needs, such as additional keys for other participants or special entry to the building outside of Music Library hours, please describe those needs here.

Additional access is requested for graduate student members of the Text Analysis Hackerspace; participating faculty member Fermín Moscoso del Prado Martín, and to-be-hired WE1S postdocs.

Security

We offer lockers in the DAHC for the storage of sensitive equipment, however the Commons is an open, low-security area, and we anticipate people circulating through the Commons during the day. If necessary, please describe additional security requirements -- such as lock bins, table security cables or equipment that should be checked out from the front desk.

Security cable for "deep learning" workstation.

Other Requirements

Any other requirements or DAHC support for the success of this project.

N/A

(D)Preliminary Budget

Funding / Expenses

Indicate anticipated expenses including both start-up and maintenance costs. Identify any other funding sources that have been secured or applied for in support of this proposal. For funds requested from DAHC, briefly outline how the requested equipment software contributes to the proposed project.

Item / Service	Cost (-)

"Deep Learning" Workstation (described under "Equipment" above)	$1,000 from the DAHC (to be supplemented by about $1,000 from WE1S Mellon funding)
Funding Source	**Contribution (+)**
DAHC	$1,000
WE1S Mellon funding	$1,000
Total Expenses:	$,2,000

Proposed Timeline

Proposals may have a total duration of up to 3 years (AY 2017-18 to AY 2019-20), aligned with the quarter system. Indicate your proposed start quarter / date to occupy the space and the duration (one quarter, one year, ongoing etc.). Are there any significant research, teaching, or event milestones that you anticipate for the project during the project period?

Three years (roughly timed to align with the duration of the WE1S project):

- Start date for Text Analysis Hackerspace: Winter 2018
- End date: Fall 2020

Major milestones will include: the summer research camps of the WE1S project in each of the summers of 2018, 2019, and 2020. Workshops for the general UCSB faculty and student community will serve as milestones during the academic years.

Submit statements of interest, questions, and completed proposals to: dahc@hfa.ucsb.edu