Supporting digital scholarship to bridge disciplinary and hierarchical boundaries

Vilja Hulden, the Digital Humanities Graduate Certificate Committee, et al.

[NB: As a demonstration of computational approaches to humanities questions, we have applied some standard tools to the set of white papers submitted before Jan 6, 2018. See the Appendix below, which contains **visualizations** and **analyses** of the words and topics in the set of white papers to provide an overview of the campus community's concerns. There is also an interactive version of the topic model visualizations available at http://bit.ly/whitepaperstopics.]

Multiple recent endeavors on campus have focused attention on the need to bring new technological approaches to the humanities and social sciences as well as humanities and social science thinking to technological realities and problems. For example, on the initiative of faculty in French and Italian, Information Science, History, and the Libraries, the Graduate School recently approved a Graduate Certificate in Digital Humanities, involving a core course in the technology and theory of digital scholarship in the humanities (and social sciences) and electives that allow students to deepen their understanding of specific approaches to, and/or explore theories of, technology. In a similar vein, the Laboratory for Interdisciplinary Statistical Analysis (LISA) has begun to run workshops and short courses on statistical training for scientists, humanists, and others who believe that quantitative approaches might be beneficial in their research. The new Center for Research Data and Digital Scholarship (CRDDS), a partnership between the Libraries and Research Computing, also runs workshops on R, programming in Python, computational text analysis, data visualization, and more; hosts regular consultations where students and faculty can bring their questions and problems; and organizes brown bag talks by faculty and students on their digital research.

As the "sold-out" workshops offered by CRDDS and LISA as well as the popularity of the talks organized by the Exploring Digital Humanities series demonstrate, the demand for training in and insight into computational approaches is intense. Several departments and institutes on campus (Information Science, Linguistics, CARTSS, IBS, etc.) are also running or planning undergraduate and graduate courses on introductory programming, data analysis, and machine learning. Graduate courses on e.g. Geographic Information Systems (in Geography) or Social Network Analysis (in Political Science) draw students from several departments and usually fill to capacity quickly. We are trying to respond to demand – for instance, creating new courses and building better infrastructure. One example is a downloadable Virtual Machine stocked with digital humanities software to make it easier to use these often experimental technologies in class (an effort that was supported by an Innovative Seed Grant). Much more, however, remains to be done.

We propose that the campus consider ways to consolidate the gains made in this area and to accelerate the momentum already achieved. Thus far much of the progress has been made due mainly to volunteer efforts by individual faculty. We would like to see funding to mold sustained

programs (with possibly additional technical support staff and e.g. improved options for web hosting). We would also encourage incentives (summer funding, course releases, etc.) to allow faculty new to digital scholarship to get involved – learning new technologies takes substantial time and is hard to accomplish in the interstices of regular teaching and research loads. We would also like to see stronger support for bridging the humanities-technology divide; existing humanities and social science research methods can be enriched by adopting computing methods to analyze and visualize the larger and more complex data sets that are increasingly available. But computing and engineering perspectives increasingly need to be sensitized to historical precedents of technological developments, cultural and linguistic differences in technology use, and the social, political, and ethical consequences of new technology, which the humanities can uniquely inform. We believe that it is essential for the future of the university and the society to not only "train" humanities and social science faculty in technology, but to foster true interdisciplinary discussions that bring scientists, engineers, social scientists, humanists, and artists together to consider the societal and cultural implications of technology.

Digital scholarship and the field commonly known as digital humanities (DH) are sites of collaboration not only across disciplines but also across hierarchical divides. Digital pedagogy emphasizes students as not only consumers but producers of knowledge, and many digital projects leverage the different types of skills produced by undergraduate and graduate students working together with junior and senior faculty and academic support staff in the libraries and in technology centers. This type of work – which many universities are now incorporating into their normal praxis and incentivizing in the form of summer grants, faculty course buy-outs, project-based classes, and more – has the potential to create productive interactions between faculty and students that advance student engagement, retention, and real-world skills as well as faculty research agendas and pedagogical expertise.

To demonstrate some common digital humanities tools, in the appendix below we provide a brief DH analysis of the white papers submitted to the Academic Futures process (see next page.)

Appendix: Analysis of white papers using common DH tools

Digital humanities tools are often used to attempt to reveal patterns in text by various "distant reading" methods – to e.g. surface topics in the texts or analyze commonly used words or collocations (co-occurring words or phrases) from the texts *en masse* rather than reading the texts individually for content.

Here we demonstrate a few DH approaches by analyzing the white papers submitted to the Academic Futures process using word clouds, topic modeling, and sentence structure analysis.

The data

All white papers submitted prior to Jan 6, 2018 (n=84), downloaded using wget and processed into text format. The total word count is 146,368 (average of 1,742 words per white paper).

This is a rather smaller corpus than is commonly analyzed using DH tools. It can nevertheless be usefully explored, even if some patterns are of course less reliable or productive than they would be in a larger body of texts.

Analysis

Word clouds of bigrams

First, the texts were combined and processed into bigrams, after removal of stopwords (common words like "the", "and", "to" etc.) (A bigram is a two-word sequence: for instance, the sentence "This is a great white paper" has the bigrams `this_is`, `is_a`, `great_white`, `white_paper`; since we have removed stopwords, the only bigrams from that sentence would be `great_white`, `white_paper`).

The bigrams were then fed into Wordle, an online service that creates word clouds (http://www.wordle.net). The larger the word (bigram), the more frequent it is in the text.

Figure 1 has a word cloud created from all the bigrams, with the exception of the bigrams "CU Boulder," "white paper" and "Colorado Boulder" (as those are so frequent they overshadow all others, and do not tell us much about the texts.)

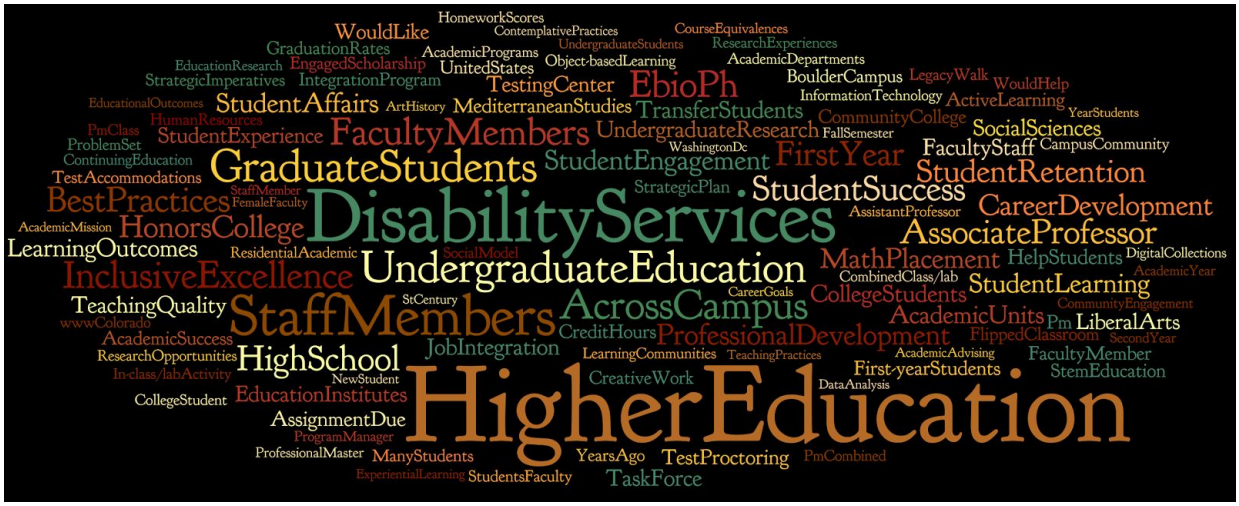

Figure 1: Word cloud of all bigrams (except CuBoulder, ColoradoBoulder, WhitePaper)

We can note the dominance of "higher education" (n=109) and, interestingly, "disability services" (n=79) and "staff members" (n=79).

If we remove some of these most common bigrams, we can drill down a little into the common-but-not-so-dominant bigrams, as shown in figure 2.

Figure 2: Word cloud of bigrams;
bigrams removed: CuBoulder, ColoradoBoulder, HigherEducation, StaffMembers, DisabilityServices

Drilling down this way highlights the focus of the white papers on students: "graduate students" (n=59), "undergraduate education" (n=50), and "student success" (n=35) are all prominent, as is "student retention" (n=32). Clearly there are other concerns as well, though: for instance, note the presence of "inclusive excellence" (n=39), "professional development" (n=30), or "academic units" (n=25). Different groups of academic disciplines are also represented: "social sciences" (n=23), "liberal arts" (n=22), and "stem education" (n=19), for instance.[1] Finally, note also the focus on innovative undergraduate teaching, including "flipped classroom" (n=27) and "undergraduate research" (n=24).

[1] The "EbioPh" is an artefact from a long list of signatures in one white paper listing "[name] EBIO Ph.D. student."

Topic modeling

Topic modeling is a technique that uses complex calculations to attempt to automatically surface "topics" from text based on word co-occurrence in texts. It relies on a bag-of-words model, i.e., the order of words does not matter. While an explanation of the technical details is beyond our scope here, it is important to note that one does not tell the topic modeler anything about the *content* of the topics; rather, one only tells it *how many* topics one wants to see, and the topic modeling software creates the topics based on co-occurring words in documents. Each document can contain several topics, and a word can belong to more than one topic.[2]

The most commonly used software package for topic modeling is MALLET (http://mallet.cs.umass.edu). To create the topic model discussed below, we ran a number of experimental topic models with MALLET, and chose one that seemed both representative and useful (as many DH methods, topic modeling is often used iteratively.) The chosen topic model infers 10 topics used in the set of white papers (each white paper is fed into the topic modeler as a separate text document, with stopwords again removed).

Figure 3 shows the topics and their key words (the words associated with each topic; they are displayed in an order of decreasing significance). It also displays a manually created label for each topic. The labels were decided upon using both the keywords of the topic and an exploration of which documents the topics were associated with (more on topic association with specific documents below).

The most common topic seems to be a rather general and possibly uninformative topic having to do with all stakeholders on campus, though it seems that this topic also addresses the need for "support" and "resources" for these stakeholders.

[2] For a more detailed introduction to topic modeling, see e.g. Matthew Jockers, "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors," 9/29/2011, http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/

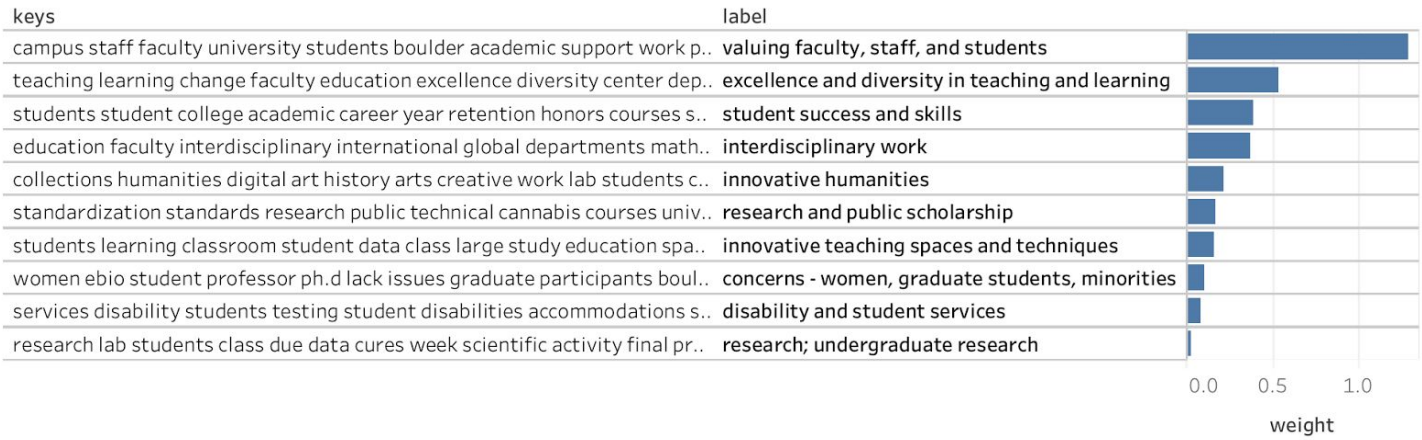

Figure 3: Topics created by MALLET from white papers

The second topic reflects the focus of the white papers on teaching and learning, and the importance of both excellence and diversity in both, while the third focuses on student success. As is clear from topics 4 and 5, interdisciplinary education and research are key foci of the white papers, as is the role of the humanities and the new digital and creative work being done in the humanities.

The topic model seems to reflect some trends in the word clouds: for example, the emphasis on student learning and success, or the importance of inclusive excellence. On the other hand, the topic model reveals that the word cloud seems to have exaggerated the presence of e.g. "disability services," as disability and accommodations here seem to represent a rather minor topic. This may be the result of using bigrams for the word clouds: perhaps "disability services" is such an established term that it gets used a lot, in comparison to other issues that are described using a greater variety of two-word sequences.

Unlike a word cloud, a topic model allows one to drill down to the level of the documents themselves. Thus, one can see which topics are prominent in which documents, and one can also visually discover patterns in topic co-occurrence. The long chart on the following pages (in Figures 4 and 5, created using Tableau) shows the distribution of the topics above in all the processed white papers (the larger the dot the weightier the topic in that white paper; note that this uses the standard Tableau view where all cells get a dot even where there is no data content in that cell, and thus the very smallest dots should be interpreted as zero.) For an interactive version (allowing e.g. sorting of white papers by topic or topics by white paper), please see the online version at Tableau Public: http://bit.ly/whitepaperstopics (the tab Story-TopicKeys has the topic definitions above, while the tab Story-TopicDistribution provides the view depicted in the images below.)

(Figure 4: Topic occurrence in white papers – continued on next page)

(continued)

(Figure 4 continued)

Drilling down into the topic occurrence in different white papers demonstrates, for instance, that interdisciplinary and international work is a concern in several white papers, and confirms that e.g. accommodation and disability issues, prominent though they were in the word cloud, appear to be of major concern mainly in a few white papers targeting those issues specifically. Innovative work in the humanities (which appears at the top of the image as the list is sorted by that topic here) is also clearly something that quite a few white papers address. The visualization also confirms that the largest topic is mainly a reflection of the common vocabulary in all these white papers, as it appears fairly prominently in nearly all of them.

Hierarchical clustering

Another way to try to extract structure from a set of texts in an unsupervised (fully automated, without first labeling some texts or guiding the process toward a particular end) manner is to try to "cluster" them. Below is a visualization of a hierarchical cluster of the white papers (performed using the statistical software package R, using the `tm` package (Ward's method, hclust). Interpreting hierarchical clusters is tricky, and drilling into this would require more time; however, it is interesting to note that there is at least some convergence between this and the topic model, as the hierarchical cluster shows one very large cluster (in green) and three relatively large ones (turquoise, purple, pink), which seems to tally with most of the 10-topic model's topic mass being in four main topics.

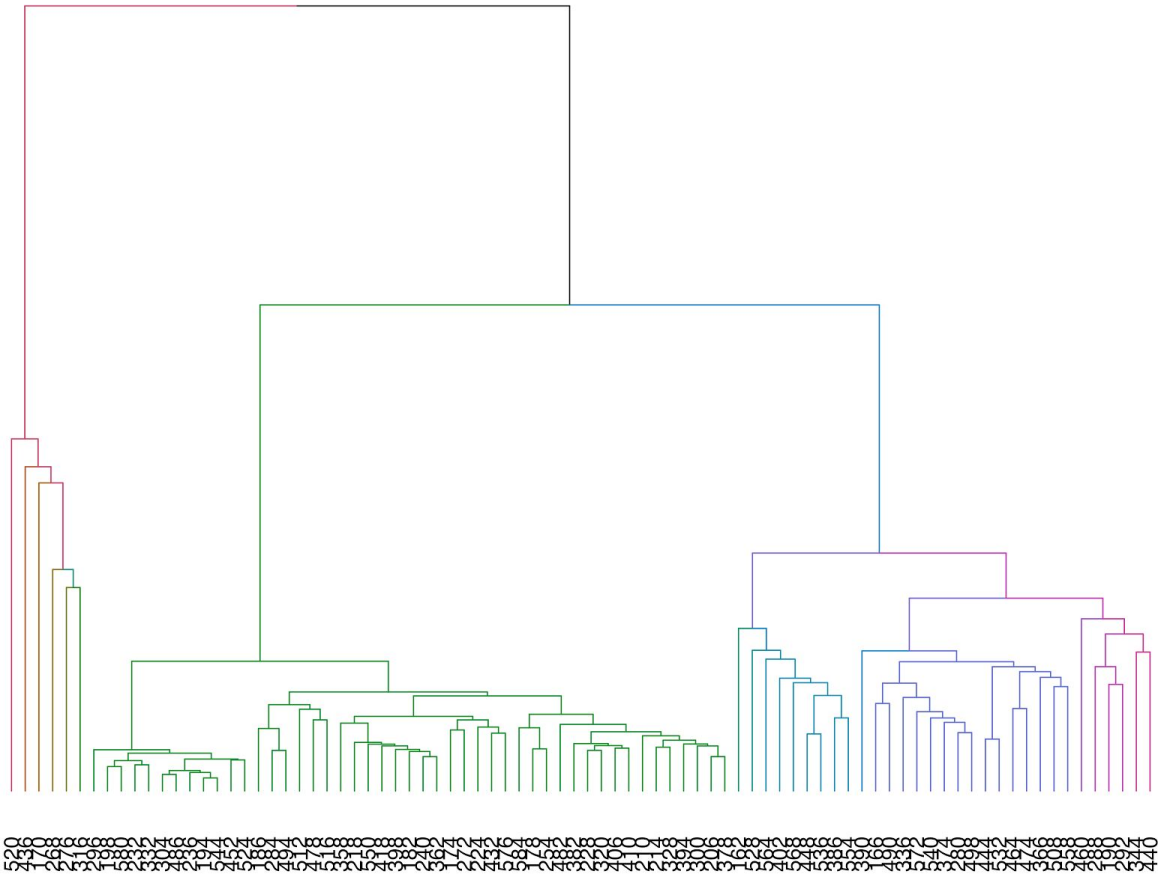

Figure 5: Hierarchical clustering of white papers (numbers at bottom are the white paper IDs)

Sentence structure analysis

Besides the bag-of-words representations above, it is of course also possible to analyze these texts in a manner that preserves and pays attention to their linguistic structure. Using the Stanford CoreNLP toolkit, all the white papers were "parsed," i.e., the part of speech (verb, adjective, etc.) of each word was marked, and the structure of the sentence analyzed. Then, using the Stanford Tregex tool, these were queried to find subjects and objects, as well as verbs and adjectives associated with particular subjects and objects. The idea here was to see if particular noun phrases appear more frequently in subject or object position, which could be read as a kind of a proxy for exercising agency or at least taking active action rather than being acted upon (though obviously this is merely a syntactic proxy for the semantic categories of *agent* [actor] and *patient* [acted-upon].) This works reasonably well with large data sets; the current data set is rather too small for this to tell us much or to be particularly reliable. To indicate the type of analyses possible, however, we report some of the results below.

First, the basic parts of speech. The most common verbs in the data (apart from *be, have, do*) are:

provide (302)	use (185)	require (157)
include (273)	create (180)	increase (146)
make (261)	help (179)	engage (141)
learn (215)	take (173)	offer (118)
need (214)	develop (170)	teach (110)
work (197)	support (167)	

The most common adjectives are:

academic (324)	own (93)	graduate (69)
other (254)	best (92)	available (69)
new (246)	different (91)	creative (68)
such (205)	large (84)	specific (67)
more (173)	major (83)	various (64)
many (155)	social (81)	better (64)
undergraduate (140)	international (81)	significant (63)
important (115)	high (81)	institutional (63)
higher (112)	public (80)	individual (63)
interdisciplinary (107)	current (80)	technical (61)
	professional (77)	particular (60)
educational (107)	diverse (74)	global (60)
first (97)	effective (70)	critical (58)
		scientific (56)

The top noun phrases by raw count are:

students (310)	student (59)	department (34)
university (105)	cu boulder (49)	group (33)
faculty (87)	staff (48)	people (31)
cu (69)	college (45)	courses (30)
campus (68)	work (42)	data (30)
research (67)	departments (35)	units (25)

As noted above, one can also try to analyze whether a particular noun phrase appears at the subject or object end of the spectrum of "subjectness" vs. "objectness." This means that it appears with relatively greater frequency in positions like "**The girl** hit the ball" rather than in positions like "The teacher praised **the girl**" or "The teacher gave the pen to **the girl**").

In this set of texts, the noun phrases that appear most strongly in such "subject" positions are:
`participants, study, students, college, student, group`

The top noun phrases appearing at the "object" end of a subjectness-objectness continuum (by the same token as above) are (with inanimate conceptual words like "information" excluded):
`institutions, program, faculty, people, staff, graduate students`

Again, what (if anything) this means is unclear given the small data set. Probably the appearance of "study" in the first list is due to "the study shows" and similar constructions; the presence of "staff" and "graduate students" in the latter list may perhaps be an indication of white papers addressing the needs of these groups ("we need to support graduate students" and similar constructions).

Finally, one could examine what verbs, for instance, go with what subjects and objects. Again, the data is rather limited for this type of analysis, and only the most common noun phrase (*student/s*) has any meaningful data on this. As subjects, students mostly *learn, take, leave, understand, report, pursue*, and *need*, while in the object position, students are mainly *provided, supported, engaged, helped, served, encouraged, assisted,* and *retained*.