4Humanities@UCSB Agenda, November 21

Topic Modeling the 4Humanities

What Every1Says Corpus: The Idea

- **Research Material: WhatEvery1Says Corpus**
- **Research Questions:** *[from agenda of 4Humanities@UCSB meeting of Nov. 7, 2013]*
    - "Our hypothesis is that digital methods can help us learn new things about how media pundits, politicians, business leaders, administrators, scholars, students, artists, and others are actually thinking about the humanities. For example, are there sub-themes beneath the familiar dominant clichés and memes? Are there hidden connections or mismatches between the "frames" (premises, metaphors, and narratives) of those arguing for and against the humanities? How do different parts of the world or different kinds of speakers compare in the way they think about the humanities? Instead of concentrating on set debates and well-worn arguments, can we exploit new approaches or surprising commonalities to advocate for the humanities in the 21st century?"
    - (Example of someone trying to rebut clichéd premises of discussions of education: Valerie Strauss, "Five Bad Education Assumptions the Media Keeps Recycling")
- **Specific research questions:**
    - What are the common "themes" (ideas, theses, evidence, metaphors, etc.) that divide or join people discussing the humanities?
    - What are the lower-level or latent themes beneath those everyone "knows"?
    - What are the outlier themes?
    - What are the patterns of connection between themes, between spokespersons, and between media outlets?
    - How do themes compare across time?
    - How are themes differentiated by nation, region, gender, age, etc.?
    - Other questions?
- **The Research Method: Topic Modeling**
    - David M. Blei, "Topic Modeling and Digital Humanities" (2012)
    - Matt Burton, "The Joy of Topic Modeling" (2013)
    - Ted Underwood
        - "Topic Modeling Made Just Simple Enough" (2012) [version annotated by A. Liu]
        - "What Kinds of "Topics" Does Topic Modeling Actually Produce?" (2012)
    - Matthew L. Jockers
        - "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors" (2011)

        - *Macroanalysis: Digital Methods and Literary History* (2013) -- Chap. 8: "Theme"
- **Other Possible Methods**
    - Corpus linguistics (e.g., using antConc) (example)
    - Social network analysis (e.g., using Gephi)
    - Information visualization (e.g., using Many Eyes)
    - [For other tools, see A. Liu's "Digital Humanities Resources"]

Early Proof of Concept

- "The Heart of the Matter" Topic-Modeled (A Preliminary Experiment)
- "The Heart of the Matter" Visualized

Strategy (1) -- Work Plan (Personnel, Time Line, etc.)

- Personnel
    - Possible plan: start project at UCSB, then ask for critique of methods and further help from international DH community.
    - Who at UCSB might want to participate?
    - Involvement of students from the "Writing and Civic Engagement" minor in UCSB's Professional Writing program. Interns could help us with workflow of analyzing the WhatEvery1Says corpus.
- Time line? Possible syncopated rhythm of 4Humanities@UCSB activities during rest of this year:
    - Discussion meetings (e.g., on "Global Humanities," "Humanities / Sciences")
    - Research workshop meetings (topic modeling, etc.)
- Lindsay will email to people who came to this meeting and we can then put together a time plan and work flow.

Strategy (2) - Methodological and Technical Workflow (known issues & draft plan)

Red = Discrete Tasks

- **Collection methodology for WhatEvery1Says corpus**

- Critically examine the selection criteria. Currently, the criteria for inclusion of resources are:
  - Online material.
  - Textual documents (not audio or video resources).
  - Documents in English.
  - Documents more-or-less in the public milieu (including journalism, blogs, reports, white papers, etc., but not scholarly studies or research).
    - There is no reason why we can't also include more scholarly works as we continue on, but we are starting with works in the public domain.
  - Documents sized between posts/articles and reports (not books).
  - Known needs:
    - Most documents in the corpus are from the last two years. We need documents from past periods. This would be a very high-value thing for us to do, even if we only had a small number.
    - Most documents in the corpus are from (or address) the North American context (with some representation of the U.K. and other nations/regions). We need more documents from other parts of the world. There are many technical, conceptual, and practical difficulties with this one: for example, what is the status of a translated document in terms of topic modeling?

- **Collection format for the corpus** (currently a Google spreadsheet). Should the holding format be (for example) a database, a Zotero collection, etc., to make it easier to filter, group, and extract metadata about the documents (e.g., citations)? **(Related issue: collection platform for processed files.)**
- **Text extraction from HTML and PDF files**. We need to research tools such as pdf2htmlEX to semi-automate the extraction of text from documents into "plain text." **(Related issue: exclusion of non-relevant material in documents such as advertisements, bios of authors, copyright notices, etc.)**
- **Text cleaning and preparation**: Python scripts or other tools for fixing common errors, standardizing spellings, resolving hyphenations, etc. (Cf. Ted Underwood and Andrew Goldstone's scripts and other resources for topic modeling).

- **Interpretive text preparation**:
    - Lemmatization: treating different forms of a word as the same thing.
    - Consolidating semantically unitary bigrams into unigrams (e.g., "social sciences" into "social_sciences").
    - Filtering out proper names (experimenting with named-entity recognizers).
    - Creating a "stop list." These are words you tell the algorithm not to consider. (See Ted Underwood and Andrew Goldstone's stop list for their project on topic modeling literary studies journals.)
    - Experimenting with parts-of-speech taggers (POS) to filter out everything but nouns (cf. Matthew Jocker's topic-modeling work).
    - "Chunking" (breaking documents if needed into appropriately sized subdocuments).
- **Topic Modeling:**
    - Experimenting to see if we should use the full-featured Mallet topic modeling suite or its Java implementation (Topic Modeling Tool).
    - Experimenting with different parameters for the topic modeling, most importantly: number of topics to ask the algorithm to produce.
    - Interpretive labeling of topics (and visualization of topics).
- Need to develop a workflow for our team, not only to make work more efficient and manageable, but also to document what we've done.

Strategy (3) – Outcomes

- Creation of interactive site for exploring the topic model of WhatEvery1Says. (Cf., DFR-Browser, a browser-based visualization interface created by Andrew Goldstone for exploring his topic model of JSTOR articles).
- Co-authored research report or article on outcomes.
- Workshop to brainstorm ways we can apply the outcomes in facilitating, guiding, or creating advocacy arguments and materials.

General Discussion

- Topic modeling tells you not just what people are talking about but *how* they are talking about it. It also has the ability to detect conversational communities, or how texts are grouped together. Do we want to identify sub-communities that are given as such, clearly characterize them as subgroups, to break up corpus even further and analyze these texts separately? OR, is one of our goals to find conversations we didn't know were there in corpus as a whole? Each approach has upsides and downsides. We can do both. The larger meta-question is about what we are trying to accomplish: Are we trying to hone in on established communities/frameworks that people already have a sense of? Are we tying to develop/make/discover new communities/frameworks?
    - We can ask different kinds of questions about discrete groups vs the entire corpus. Then we can perform social network analysis on these groups to discover what groups aren't talking to each other, how we might get them to talk to one another, etc.
- Wouldn't it be better to work with an already established corpus so that our results are better? So that results aren't determined by randomness of selection of things to go in corpus.
    - No reason why we can't do both. There are limitations on pre-existing corpuses that are out there.
- Another thing we could also do would be to establish rules for inclusion in the corpus that would better define what the corpus is.
    - Could also consider pre-existing plain text corpuses of sites like Facebook, Wordpress. But the problem with these is that they are generally pre-digested.
- Ultimately what we want to get out of this is "What are priority arguments?"
    - Another thing we want is to get a sense of how discussion has evolved over time.
- One place to look would be news corpuses, or things like Gutenberg.
- What about TV transcripts from news shows?

    - Mandated closed-captioning in certain countries. We want to find people who have already collected these kinds of transcripts and use what they have collected.
- Sensible organization plan:
    - Working group 1: Collection methodology
    - Working group 2: Text preparation (including interpretative and technical parts of text prep)
    - Lindsay will email everyone in attendance at meeting today to ask about working group preferences.
    - When we reconvene at beginning of next quarter, we will have a sharper sense of what needs to be done as far as workflow and planning goes.
        - Hope is that by the end of the academic year, we will have enough preliminary results for Alan to take to the international community and get their feedback.
- Priscilla's idea about including Honors students, students would receive an Honors credits for working on the project (students need 2/year to stay in the Honors Program). This would interest a lot of students. Alan will check on this.
- We need a public workplace where everyone can put their materials and people can tell that things are happening, etc. The amount of energy required to coordinate things is reduced; can always be making progress if there is a place where all of our documents are shared.
    - Alan will set up PBWiki for this purpose.