Reading Week 04 - Overview of Computational Text Analysis

Add your name here

June 16, 2021

Abstract

This week's readings are two overviews of Computational Text Analysis. The first paper is written by an interdisciplinary group of computer scientists and social scientists. The second paper is written by a Professor of Computational Social Science at London School of Economics

1 Instructions

For each reading, include the following:

- 3-4 sentence summary
- 1 sentence about something in particular that you like
- 1 sentence about something you didn't like or something you found confusing and you'd like me to explain
- 1 question for future work or one sentence on how you'd apply the reading for a final project

Then, in section 4, write a small comparison about the two papers based on the instructions.

2 How We Do Things With Words: Analyzing Text as Social and Cultural Data

https://arxiv.org/pdf/1907.01468.pdf [?]

This article focuses on breaking down the methodology behind computational text analysis within a social application. Rather than looking at the text as a purely qualitative source, which must be read linearly and thoroughly, the article questions the epistemology of textual interpretation and how knowledge derivation extends not only to a "classical" reading of a text but also to a properly computed analysis of textual information. Similar to how numbers can be read and understood, but also computed, textual data can be exploited via analyzing its various properties in relation to quantity, variation, and categorization. Nguyen et. al speak concisely about the process from acquisition to interpretation, and the authors make explicit claims to the potential issues and boundaries to computational applications while still endorsing the larger usefulness of the methodology. I found it important that the authors fore-fronted larger epistemological and implementation-oriented issues of computational textual analysis, and they explicitly stated that "operationalizations are never perfect translations, and are often refined over the course of an investigation, but they are crucial" (2).

Similar to the origin of statistics, where one must maintain a critical in the application of the mathematical calculation, users of computational text analysis must always remember that the computed outcomes are simply approximations to real phenomena. I still have many questions in relation to the "Born-digital data" subsection; not in regards to what the authors are explicitly saying, but in relation to the possible ethical issues that they are raising. Who gets to become the overseer of digital equity and fair use, and how can we ensure that certain kinds of data doesn't get exploited for personal use by third parties with ill intentions? In a world with an abundance of digitally-produced information and seemingly little regulation, it appears that one must simply expect the exploitation of his or her produced text. I think this text would be good for thinking about my final project with regards to how one should iterate through the process of exploiting textual information via computational methods, as well as how can I confirm my findings.

3 Text as Data: An Overview

https://kenbenoit.net/pdfs/CURINI_FRANZESE_Ch26.pdf []

Throughout this article, Benoit attempts to look at the concept of "text as data" in a way that illuminates the complexities in attempting to view textual products as an exploitable form of data. Benoit differentiates between the concepts of "text as text" and "text as data" by explaining how "all forms of text contain information that could be treated as a form of *data*" but that text only begins to become data "when we record it for reference analysis" (463). In order to transform the text, there needs to be some sort of abstraction or use of structuring in order to turn the textual information into operable and quantifiable chunks of info. The author goes into explaining much of the linguistic features embedded in any piece of text, and how these internal structures can be exploited for external purposes. What I find fascinating is that just two centuries ago (and in many ways, even to this day) statisticians argued about the self-evident nature of numbers and quantitative data, and we now understand that there is a more complex relationship between numbers and the world they claim to represent; even though Structuralism only recently provided the framework through which we could understand language in a object-oriented way via linguistics, computational linguists have been able to exploit the massive computational possibilities of text via their inherent internal structures. Much like the many arguments surrounding statistics, there seems to be a large and hearty discussion about the nature of treating "text as data" as a quantifiable and epistemologically-just methodology — what are the possible consequences of treating human textual products as some kind of computable object?

I wrote many notes in the margins related to different linguistic and translation studies texts that I have read, and I have some questions about modern theories of interpretation of text (i.e. hermeneutics, New Criticism, etc.) play a role in our understanding of formalist approaches to textual interpretation. Though I would argue that only because of the rise formalism that there seems to be a wider use of such methodologies, the advancement of computational text analysis has even begun to allow for para-textual analyses of tangential corpora to specific bodies of text. I would like to apply this reading to my final project by thinking more about how I can be explicit in what features I wish to extract in order to properly transform text into data via a defined methodology (thus to arise some awareness to a latent characteristics).

4 Comparison & Contrast

Instructions: Write about 3 sentences comparing the two readings. What was similar across both and what seems to differ

As foreshadowed in the "How We Do Things With Words" article (shortened to HWDTWW from hereon), there is an undertone within computational text analysis that it is inherently understood that language must be treated uniquely; however, the "Text as Data" article makes further clarifications to how and what different handling of text means. Both articles speak about different methods of conducting computational analyses of text, but "Text as Data" leans towards a linguistic side of things, whereas "HWDTWW" leans towards a computational/procedural interpretation. While "Text as Data" gives explicit complications to what makes text into data and explaining the processes therein, "HWDTWW" focuses on the concepts on a procedural level, moving from beginning to end in steps. Both texts were useful in understanding the wide-array of developments in computational text analysis, and both displayed an extended framework of scholarship that shed light on the larger discourse currently in process.