Corpus-building for Computational Text Analysis

Taking it Further with Corpus Building

This handout shares more information on how you could build and analyze a corpus of your own.

Key Words

- **Computational Text Analysis**: Text analysis is making inferences based on textual data. Computational text analysis (CTA) involves a computer drawing out patterns in a text, and a researcher interpreting those patterns. CTA includes methods such as word count frequency, nGrams, and sentiment analysis. CTA is similar to statistical analysis, but the data are texts.
- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stopwords include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.

How to Build a Corpus

When building a corpus, especially in the context of a smaller project, follow these steps:

1. Choose the texts you would like to include in your corpus.
    - Remember that the specific texts you select will have a significant impact on your results, and think carefully about how you can find texts that are as representative as possible of the full set of potential texts.
    - Carefully define the scope of your corpus (for example, by publication date or location, genre, author, etc.).
    - In any analysis of your results, you should address the contexts of these texts, and consider any possible limitations in their ability to serve as proxies for the phenomena you wish to study.
2. Create a folder on your computer or cloud storage where you will store your corpus. Make sure not to put any spaces in the name of your folder.

Developed by: Sarah Connell, Colleen Nugent, Margarida Rodrigues, and Yunus Emre Tapan
Questions? Contact us: nulab.info@gmail.com

3. Once you have chosen the texts you will include, open a plain text editor (for example, Notepad on PCs and TextEdit on Macs).
    - The individual plain text (.txt) files that will make up your corpus are machine readable versions of the documents you chose to include in your corpus. "Plain text" just means that the files don't have any hidden formatting.
    - TextEdit on Macs: You must make sure it is configured to work with plain text files. To do this, open TextEdit and go to “Preferences” and make sure “plain text editor” is selected. Then, restart TextEdit.
4. Creating plain text files:
    - Usually, you can just copy-paste from your source files into each .txt file. It is also worth investigating whether your texts are available to download directly as .txt. If so, you can download these and save them in the folder you created instead.
    - **Only copy one text** into each new plain text file (unless you are combining texts from similar resources for research purposes).
    - Some articles might have HTML/web-browser versions that will be easier to copy-paste than PDFs. Some PDFs will be copy-pastable and others won't. If you want to work with files that cannot be copy-pasted, you might need to transcribe them. Or, if you're feeling adventurous, you can look into Optical Character Recognition (OCR), which attempts to automatically transform PDFs into text. There are several free tools for running OCR on files that you can find online.
    - Make sure each file name ends with .txt
    - Use filenames to indicate the data inside (ex: “browne_chapter-1.txt”)
    - Make sure not to put any spaces in the names of the files as you save them. Use underscores or hyphens to mark spaces between words instead.

Where to Find Texts

You can start by browsing NULab's list of datasets, which contains links to repositories of texts, including Project Gutenberg—a repository of over 60,000 free eBooks in plain text—and Northeastern’s Early Caribbean Digital Archive (ECDA), containing pre-twentieth century Caribbean texts, maps, and images. Another resource for finding datasets is Kaggle.

Web-Browser Computational Text Analysis Tools

These browser-based GUI (Graphical User Interface) text analysis tools can show word frequencies and patterns in language. While using coding languages like Python and R can open up other types of analysis (such as word embedding models and topic modeling),

these GUI tools allow you to do more basic analysis to begin examining your texts computationally.

- **Word Trees**: This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest. There are some restrictions in size; fewer than 1 million words should work, but loading that much text in might be slow. http://jasondavies.com/wordtree/
- **Word Counter**: This is a user-friendly basic word counting tool; it allows you to count single words, bigrams, and trigrams in plain text files and to download spreadsheets with your results. The max file upload is 10MB. http://databasic.io/en/wordcounter/
- **Lexos**: This is a tool for preparing and analyzing digital texts; it offers several options for text preparation, and a wide range of different analytical possibilities. Importantly, Lexos also preserves all the changes that are made to a text, so that any results can be reproduced. http://lexos.wheatoncollege.edu/
- **Voyant**: This suite of tools gives you counts of words and lets you compare patterns in word locations and frequencies, or examine keywords in context, along with a few other options. Voyant will let you upload larger files than most other interfaces (up to as many as 4 million words, though it may take more than one try to successfully upload very large files). http://voyant-tools.org/

Other Computational Text Analysis Tools

These are more advanced tools that you can explore.

- **AntConc**: A corpus analysis toolkit for concordancing and text analysis. AntConc is free and available to download for MacOS and Windows. AntConc performs better with many small files, rather than one or two large ones—there is no limit on how many words you can analyze, but larger corpora will take longer to work with https://www.laurenceanthony.net/software/antconc/.
    - You can find a useful tutorial on AntConc by the Programming Historian, where you can also find a range of other tutorials on analyzing texts. Another helpful AntConc tutorial is made available by Manhattan College.
    - AntConc offers more options than the web-based tools above but doesn't require any command-line coding.
- If you are looking for more advanced tools and lessons, check out the NULab text analysis resources and this handout.

Northeastern-specific resources

The resources below are available through the Northeastern University Library.

- **Constellate:** This is a platform for learning and performing text analysis, with both code notebooks and a web-based interface. A free JSTOR account is needed for

Northeastern users to log in to Constellate. To access Constellate, go to the Library's list of databases and log in with your Northeastern credentials. https://constellate.org/

- **ProQuest TDM Studio**: This resource provides both web-based and code-based options for exploring textual analyses with ProQuest datasets. To get started, create an account with your Northeastern email. https://tdmstudio.proquest.com/