A Brief Introduction to Computer-Aided Text Analysis

A Resource for Public Service Professionals

Important information is increasingly being shared in textual digital format. In public service fields, examples include text alerts and social media posts from municipal agencies, constituent service requests, call center transcripts, and claims submitted to municipal risk funds. Although rich in information, textual data like this can pose a challenge to analysts who wish to better understand and solve civic challenges. This primer gives a quick overview of using computers to analyze big chunks of text, which we call computer-aided text analysis (CATA). When possible, we provide links to additional information and resources that you can use to further develop your knowledge and get started with basic analyses.

What is Computer-Aided Text Analysis Useful For?

While an individual person can interpret texts when they are small in number, CATA techniques are used when the amount of language that needs to be analyzed is too large for individual people, or even teams of people, to handle on their own. For example, one research team used text analysis on 40,000+ Facebook posts to study local governments' social media communication with citizens.[1] Another study used text analysis techniques on 281,000 reports about unhoused individuals to a social services app to better direct homelessness outreach.[2] In both cases, analysts used computers to efficiently extract and derive meaning from massive amounts of text, leading to better understanding of municipal processes and improved outcomes.

Different techniques are useful for performing different types of tasks and answering different questions. Here are several commonly used approaches to analyzing large amounts of text:

Approach	Description	Example
Quantification	Quantitative content analysis techniques break a document down into units, such as words or phrases, and count them. It is useful for measuring how often and under what circumstances certain types of words or phrases are used.	Haeder and Web Yackee (2015) used plagiarism detection software to measure the extent to which the text in draft regulations changed after interest group lobbying.
Classification	The goal of classification is to group texts into categories according to shared characteristics in their structure or content. The categories can be defined by the researcher before analyzing the data, or discovered through the process of analysis.	ten Veldhuis et al (2013) used citizens' calls about drainage problems to municipal call centers to classify failure mechanisms behind flooding issues. This helped to quantify and predict flood risks in urban areas.

[1] Hofmann, S., Beverungen, D., Räckers, M., & Becker, J. (2013). What makes local governments' online communications successful? Insights from a multi-method analysis of Facebook. *Government Information Quarterly*, *30*(4), 387-396. [subscription required]

[2] Wilde, H., Chen, L. L., Nguyen, A., Kimpel, Z., Sidgwick, J., De Unanue, A., ... & Vollmer, S. (2021). A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data & Policy*, *3*, E2. doi:10.1017/dap.2020.23 [open access]

Approach	Description	Example
Scale and indices creation	This technique uses text to create scales or indices that measure qualities. Often these qualities are hard to measure directly (latent).	Pandey, Pandey, and Miller (2017) created a school district 'innovativeness' scale derived from the content of letters to the New Jersey Board of Education written by district administrators.
Sentiment analysis	A widely used technique that gauges the feelings in text. It rates words based on positivity or negativity, then adds up the scores to measure overall sentiment. It can also detect specific emotions like anger or happiness.	Zavattaro et al (2015) measure the sentiment of city government Twitter posts and find that a positive tone, as opposed to a neutral/informative tone, encourages citizen interaction with city social media accounts.

Key CATA Terms

The following terms are used frequently among those who use computer aided text analysis techniques.

Natural Language

Natural languages are languages people use to communicate with each other (such as English). Natural languages develop organically through the course of human interaction. They differ from constructed programming languages (such as HTML), which are used to communicate with computers and are constructed purposefully for that task.

Corpus

A corpus is a collection of texts written in a natural language. These collections could come from books, webpages, reports, social media posts, or newspaper articles, to name a few examples. A corpus serves as the raw data for a text analysis.

Unstructured Data

The data found in a corpus is unstructured. While structured data takes the form of clearly defined rows and columns (think of tables in a database), unstructured data is not organized into a structured, predefined format. This makes unstructured data more difficult to work with systematically.

Information Extraction

Information Extraction is the process of turning unstructured text data into structured information that computers can work with more easily. Once in a more structured format, computers can analyze the text much like numerical data is analyzed, by quantifying or applying algorithms.

Natural Language Processing

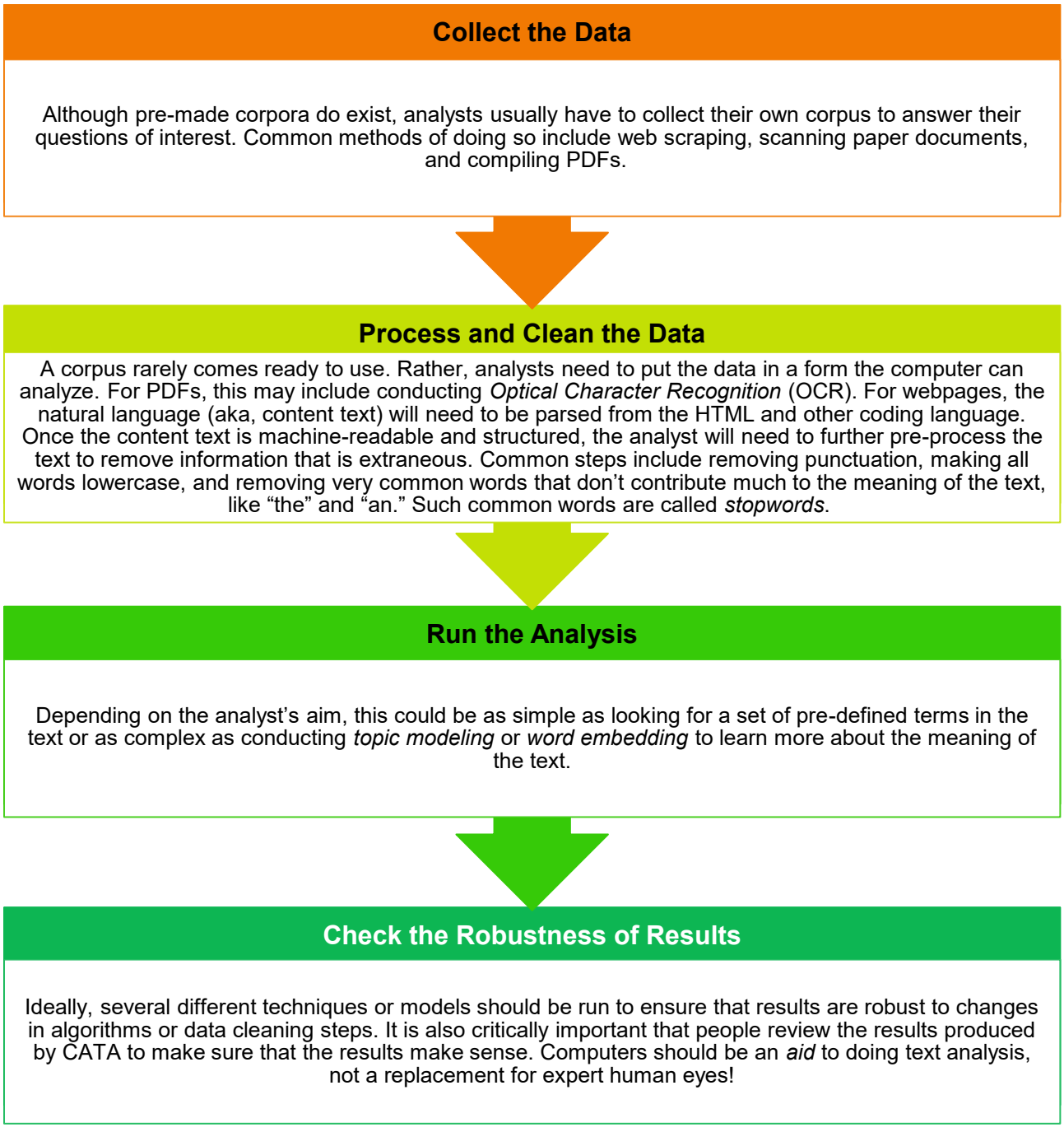
Natural Language Processing (NLP) is a technology in which computers *interpret* written or spoken human language. NLP drives many of the features we enjoy in modern communication technologies, such as chatbots, recommendation systems, and automated transcription.

Supervised and Unsupervised

In NLP, supervised and unsupervised approaches are different ways of enabling the computer to learn from the data (aka, machine learning). In supervised approaches, the analyst provides the computer a subset of the data that is labeled by humans so that the computer can learn how to correctly classify or predict the labels of the rest of the data. In unsupervised approaches, analysts apply algorithms to large amounts of unlabeled data so the computer can learn and uncover the patterns in the data.

Typical Steps in CATA

Although there are many different approaches and techniques for conducting computer-aided text analysis, they all generally require analysts to take the following high-level steps:



Skills Needed

Python and R are the programming languages considered ideal for conducting natural language processing. Luckily, software for working in these languages is free and many user-written R and Python packages are designed specifically for doing CATA. If you already know one of these languages, then you are well on your way to being able to get started doing text analysis. If you do not know one of these languages, see the list of resources below for getting started in learning. For less computationally complex text analysis techniques, programming in R or Python may not be required. We've also listed below some digital tools you can use to quantify or classify text without needing to program.

Useful Resources

There are numerous guides and tools available for doing text analysis. Below we list a few to get you started, with an emphasis on tools for those who don't know how to program in R or Python. For specific recommendations about R and Python packages to use, see the UCDavis Data Lab (https://datalab.ucdavis.edu/text-and-nlp/).

Classes on Python, R, Machine Learning, Text Analysis		
Coursera	Free	online classes in Python and R
Kaggle	Free	online classes on Python, Machine Learning, and other data science topics
LinkedIn Learning	Cost varies	offers courses in Python and R for a fee and allows participants to post a completion certificate on their LinkedIn profile
ICPSR Summer Program	Cost varies	quick introductory courses on R, Python, and Text Analysis ranging from a few days to three weeks long. Classes are taught in person or remotely by expert instructors.

Where to find data	
Public use text and audio data from user communities	Kaggle (https://www.kaggle.com/datasets) Hugging Face (https://huggingface.co/)
Secondary data collected by researchers in various formats, including text	openICPSR (https://www.openicpsr.org/openicpsr/) Harvard Dataverse (https://dataverse.harvard.edu/)

Resources: Convert images/pdfs to text
Many corpora you collect yourself will consist of scanned images or PDF files. To convert a small corpora of image files to PDFs and PDFs to text files without using R or Python, you can use extensions in Google Docs or Adobe Acrobat.

Resources: Run Analysis	
To quantify differences across versions of texts, use open-source plagiarism detection software	One example: WCopySource (https://plagiarism.bloomfieldmedia.com/software/wcopyfind/)
There are stand-alone text analysis software for those who don't know how to program in R or Python	The Linguistic Inquiry and Word Count (LIWC) software (https://www.liwc.app/). Users should evaluate how well the default LIWC word dictionaries reflect the context of their data, however.
Some qualitative data analysis software can facilitate quantitative content analysis, with the added benefit of being able to point and click through analyses.	For instance, MAXQDA (https://www.maxqda.com/) allows researchers to calculate descriptive statistics of word use, create dictionaries, and do text data visualizations.
Even if you don't know how to program in Python, you might be able to follow step-by-step tutorials	One example: Hugging Face tutorial on how to do sentiment analysis: https://huggingface.co/blog/sentiment-analysis-python

Acknowledgements

In addition to the references listed below, this guide draws from information provided at the ICPSR Summer Program 2023 *Data Science and Text Analysis* course taught by Caleb Pomeroy and Kelsey Shoub. Several examples of NLP use in public administration research were discovered in Hollibaugh (2019).

References

Grimmer, J., Roberts, M. E., & Stewart, B.B. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Haeder, S. F., & Yackee, S. W. (2015). Influence and the administrative process: Lobbying the US president's office of management and budget. *American Political Science Review*, *109*(3), 507-522. [subscription required]

Hofmann, S., Beverungen, D., Räckers, M., & Becker, J. (2013). What makes local governments' online communications successful? Insights from a multi-method analysis of Facebook. *Government Information Quarterly*, *30*(4), 387-396. [subscription required]

Hollibaugh, G. E. (2019). The use of text as data methods in public administration: A review and an application to agency priorities. *Journal of Public Administration Research and Theory*, *29*(3), 474-490. [subscription required]

ten Veldhuis, J. A. E., Harder, R. C., & Loog, M. (2013). Automatic classification of municipal call data to support quantitative risk analysis of urban drainage systems. *Structure and Infrastructure Engineering*, *9*(2), 141-150. [subscription required]

UC Davis DataLab (n.d.) Text and NLP Toolkit. Retrieved from https://datalab.ucdavis.edu/text-and-nlp/

Wilde, H., Chen, L. L., Nguyen, A., Kimpel, Z., Sidgwick, J., De Unanue, A., ... & Vollmer, S. (2021). A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data & Policy*, *3*, E2. doi:10.1017/dap.2020.23 [open access]

Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A sentiment analysis of US local government tweets: The connection between tone and citizen involvement. *Government Information Quarterly*, *32*, 333-341. [subscription required]

This material is based upon work supported by the National Science Foundation under Grant No.1829724. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation