
Automated Text Categorization and Metabarcoding Reinforces *Cryptococcus Neoformans*–Woody Decomposition Association

DeAndre Tomlinson

Publication Date

21-12-2023

License

This work is made available under a Public Domain Mark 1.0 (No Copyright) license and should only be used in accordance with that license.

Citation for this work (American Psychological Association 7th edition)

Tomlinson, D. (2020). *Automated Text Categorization and Metabarcoding Reinforces *Cryptococcus Neoformans*–Woody Decomposition Association* (Version 1). University of Notre Dame.
<https://doi.org/10.7274/24884850.v1>

This work was downloaded from CurateND, the University of Notre Dame's institutional repository.

For more information about this work, to report or an issue, or to preserve and share your original work, please contact the CurateND team for assistance at curate@nd.edu.

As a frequent library user since freshman year and an employee in the library system, I had the opportunity to experience the library from a variety of perspectives. The most long-lasting impression I will have from the Hesburgh Library system will be from my research project, which was largely powered by individuals and resources from the library. I am a first author on the pre-published research article tentatively titled “Automated Text Categorization and Metabarcoding Reinforces *Cryptococcus Neoformans*–Woody Decomposition Association” with the Michael Pfreder Lab in the Department Biological Sciences.

The timeline for this project has two parts, both of which involve staff and resources from the Navari Family Center of Digital Scholarship, a subgroup within the Hesburgh Library System. My research mentor from the Pfreder lab and member within the Center for Digital Scholarship David Molik formulated the initial idea and pipeline for this research project. *Cryptococcus Neoformans* is a fungal pathogen that kills approximately 180,00 individuals each year, with immunocompromised individuals in sub-Saharan Africa being most of the fatalities. The common approach to this problem in the research and medicinal field is to determine the pathology of the disease; however, relatively little is unknown about the ecological environments it manifests in. Therefore, there was an opportunity to learn more about this neglected disease and add to the scientific wealth of knowledge to combat and mitigate the spread of this disease by learning more about the environmental patterns and location niches of *C. neoformans*. The strategy for this project was to collect journal articles that mention *C. neoformans* to determine what is similar between all the papers and hopefully find a theme between them. David and a graduated lab member created a script to scrape Sequence Read Archive (SRA) biological data across the internet, which is essentially the DNA fingerprint of *C. neoformans*. By finding the SRA datasets that contain *C. neoformans*, they were able to find the journal article that mentioned it explicitly as well. This was possible because Hesburgh Library has access to a wide array of publisher websites and academic subscription services, like Wiley, Oxford, JSTOR, Elsevier, Nature, and more. They collected these papers and aggregated them to find the similarities between them. Originally, the process for determining similarity was powered by Non-negative Matrix Factorization (NMF) with help from Eric Morgan, another member in the Center for Digital Scholarship. This approach uses matrix multiplication and clustering to determine the most representative words and topics from the documents. However, the results from this approach were inconclusive and unsatisfactory for the scope of the project, which led to an impasse.

When I decided to take on this obstacle, I came with a unique approach that the group did not previously consider. In my Data Science class with Professor Meng Jiang, my final project for the class revolved around natural language processing. I decided to combine my project with David and the Center for Digital Scholarship and restructure the approach for tackling this project. Firstly, more documents were added that did not have *C. neoformans* for comparison and later analysis. These papers were organized and collected with the help of Natalie Meyers, another member of the Center for Digital Scholarship. She was responsible for introducing Zotero, a library tool that allows for easy management and citation creation for a variety of different documents and text types. In a research paper containing hundreds of other documents that need to be cited, she helped in the organization and citation management aspect for this project. David integrated the current progress that was made into an Open Science Framework (OSF) website, which is an open-source project management tool that supports researchers and

acts as an easily accessible collaboration tool. Because of this tool, we were able to combine our various Google Drive, GitHub, and Mendeley accounts into one environment. For team projects with multiple members editing multiple documents, this was a great tool to track progress and maintain organization.

In addition to the organizational and technical assistance from the Center of Digital Scholarship, I also utilized the public resources at the main Hesburgh library. On multiple occasions, I reserved a multimedia room so the lab or group members could meet and discuss the project with the TV screens and private sound resistant rooms. In fact, prior to the closing of campus due to COVID-19, we had an all-nighter in the Center for Digital Scholarship to write the rough draft of the research project. The 24-hour public areas and extensive technological resources were greatly beneficial towards enabling long hours for research and journal writing.

I am unable to explain at length about the research process we used to complete this project, since the article is unreleased, but I can explain the general workflow of the project. We believe with mathematical certainty that there is a connection between *C. neoformans* and woodland environments, which is reinforced by other independent research labs uncovering similar data. I used a newly created modeling algorithm and a machine learning classifier to create models and evaluate them for statistical significance. With the results of this research project, we intend to introduce the field of ecology and tropical diseases to a new algorithm that can be used for a variety of other purposes. This includes, but is not limited to, topic analysis, text analysis, thematic discovery, large text corpus, and big data parsing.

The Center for Digital Scholarship and the Hesburgh Library have been invaluable to the creation and consistent development of this project. Multiple staff members will be co-authors on this project, which shows the value that the institute brought to this endeavor. The new tools, such as Zotero and OSF, helped the workflow immensely and enabled the group to function with simplicity and precision. I am grateful for the opportunity to learn and grow as not only a student, but also a driven researcher with the help of Hesburgh Library services during my time at the University of Notre Dame.