Functional Approach to Library of Babel for Text Analysis

Dakota Sanders, *Graduate Student, University of Tennessee,* Zhixiu Lu, *Graduate Student, University of Tennessee,* Bradley White, *Undergraduate Student, University of Tennessee,* and Rachel Offutt, *Graduate Student, University of Tennessee*

***Abstract*—The Library of Babel is a permutation of every possible combination of characters as words, on every possible combination of pages, in every possible combination of 410 page books. Currently, any useful data is drowned out by nonsensical words acting as "noise." We are proposing potential objectives, depending on the feasibility of of initial mathematical investigation: either a functional implementation of the Library of Babel or a tool for effective and concise data parsing to make the current Library of Babel more navigable and usable.**

***Index Terms*—digital, archaeology, library, Babel, implementation, algorithms.**

I. Background

THE Library of Babel was first proposed as a thought experiment by Jorge Luis Borges in his short essay titled "La Bibilioteca Total" in 1939. The concept is quite simple: what if there existed a library, with a combination of every possible combination of letters, and every combination of all possible words, and thus containing all works of literature, past, present and future? In his essay, Borges muses that the library would contain detailed records of historical events passed, and even contain events of the future in perfect accuracy. Unfortunately, this library would also contain combinations of letters that are not recognized by any language as words, and thus the infinite knowledge contained in this theoretical library is perpetually drowned out by the noise of nonexistent words and nonsensical mismatches of words.

In 2015, Jonathan Basile was inspired by the works of Borges and committed to creating a digital version of the Library of Babel. In his implementation described by Borges, the Library is organized as the following:

> "The universe (which others call the Library) is composed of an indefinite, perhaps infinite number of hexagonal galleries…The arrangement of the galleries is always the same: Twenty bookshelves, five to each side, line four of the hexagon's six sides…each bookshelf holds thirty-two books identical in format; each book contains four hundred ten pages; each page, forty lines; each line, approximately eighty black letters."

Thus, we have the way to organize the data. For a naming convention in terms of finding a specific section of data, take for example: jeb0110jlb-w2-s4-v16. jeb0110jlb-w2-s4-v16 means the book you are reading is the 16th volume (v16) on the fourth shelf (s4) of the second wall (w2) of hexagon jeb0110jlb.

In the current version of the digital Library of Babel, the library claims to contain every possible page of 3200 characters, but not every possible book containing every possible permutation of these pages, due to size and time constraints.

II. Project Details

A. Objective and Motivation

Our motivation for our project stems from our innate interest in the randomness of the Library of Babel and the structure contained within. We seek to quantify at what point the random assemblage of both language recognized and unrecognized words become structured, and if there is a way to quantify the amount of expected structure given a random dataset of this magnitude.

Our objective is simple conceptually but more difficult to actually implement: can we make a tool to analyze the current web version of the Library of Babel if the current web version is feasible for web scraping? Due to the way pages are served and structured, this may not be possible. If our initial investigation into the mathematical implementation and nature of the way the Library of Babel website is structures proves that the website is unsuited to web scraping, we plan on developing a working subset of the Library of Babel that is more parsable and suited to textual analysis while still holding true to the tenets of randomness which the Library of Babel is founded upon and then creating an analysis tool.

B. Data Acquisition

First and foremost, before any data is acquired, the mathematical feasibility of implementing a functional compact version of the Library of Babel needs to be addressed to determine what data is necessary. For the most part though, data will be scraped or modeled off of the digital archive of the Library of Babel, https://libraryofbabel.info.

III. Workflow and Expected Outcomes

A. Workflow Assignments

Our team is comprised of four individuals: Zhixiu Lu, Cody Sanders, Rachel Offutt, and Bradley White. The team will approach the analysis portion of the project fairly equally, as we all need to be aware of the mathematical underpinnings prevalent behind the implementation of the Library of Babel;

Special thanks to Preston Provins and David Kennard, the teaching assistants of this course.

Special thanks to Dr. Audris Mockus, instructor of this course.

Manuscript received September 25, 2019.

however, after the analysis is completed and we determine whether or not the current implementation of the Library of Babel is feasible for web scraping and textual analysis, specific roles will emerge for each member. We will need team members to scrape data from the web if re-implementing the Library of Babel is possible or to create an outward facing API made directly for textual analysis otherwise. If we determine from our analysis that using the actual website in not feasible, we will then focus our efforts on creating a more easily parsable subset of the Library of Babel that still maintains the integrity of the original concept.

From there, we plan on having our team break into two subsects to answer the following questions for textual analysis: can we quantify the amount of randomness in a given subset of this this magnitude and is there a possible way to predict a measure of order from the chaos of the noise in the Library of Babel generated by the non-language recognized words.

B. Timeline

As of the date of this proposal, there are approximately two and a half months left in the semester to complete our work. As the details of our project implementation will shift and change throughout the upcoming weeks, the following timeline estimates are not concrete dates. These milestones are expected to change often, both in scope and expected time, and this document should not be used as reference material in this aspect.

The initial steps of our workflow revolve around a thorough investigation of the current Library of Babel site, https://libraryofbabel.info. As we are not certain of the details of this implementation or the usability of such a site, this will require significant investigation. This study will be accompanied by a mathematical analysis to see the feasibility of implementing the Library of Babel in a functional way. We anticipate the current website implementation to not work exceptionally well with scraping projects, but suspect that a subset of the contents of the Library would suit such a need well enough to remain functional for general analysis. The investigation into these details is expected to take approximately two weeks, concluding at the beginning of October 2019.

With the investigation complete, we aspire to begin a concrete development cycle to implement either a functional variant of the Library of Babel, or a useful analysis tool for a current implementation of it. By utilizing project management tools found within Github and other related software, we anticipate a lightweight development flow that is agile enough to shift and change based on the results of our work. Recent industry understanding has found, however, that this Agile workflow tends to focus on development more than delivering by deadlines. To avoid this flaw, at the beginning of this development process we plan to set out major milestones decided upon by the results of our investigation, and update these milestones as we continue development. The bulk of this development should take the entirety of the month of October, with the goal to be near completion by the middle of November.

This deadline leaves little time left in the semester, but we anticipate that the time remaining is enough time to complete an analysis of the tool we create, as well as enough time to generate documentation and presentation details to share our work with the class. This work is anticipated to complete at the end of the semester.

C. Expected Outcomes

With such a broad topic of interest, the expected outcomes of our work is yet to be decided concretely. While there are some outcomes that surely must come out of this in order to proceed, other goals are ephemeral and are subject to heavy modification or pruning.

Of the concrete goals we already have, the foremost attempt must be analysis of both the digital archive of the Library of Babel, as well as a mathematical analysis of the concept of the Library, and a discussion as to whether a proper subset of the Library is a useful addition to our work. This analysis is paramount if our work is to continue, as we are starting with a very general understanding of the system. Once we more-fully comprehend the scale of the Library, we can narrow down requirements for a useful and effective development cycle.

Following this analysis, we will decide on a path for our project. As we understand the issue presently, there are two major options our work could attempt. Option one involves reimplementing the Library of Babel archive in a way that is more conducive to data analysis and web scraping. The current application found on the web does not lend itself to being scraped very easily, primarily due to the sheer size of the corpus. Because of this limitation, the scope of our analysis may potentially be restricted. As such, a reimplementation of the archive would provide us with an easily-scrapable version of the text that would then precede much more robust analysis. Option two relies on the current implementation of the Library, and involves focusing heavily on the analysis of this corpus. While briefly introduced in Section II, our analysis aims to understand whether it is possible to quantify the amount of structure found within a randomly generated corpus of a given size. We would like to formalize this question and determine the best way to quantify such an abstract concept.

If option one is attempted, then the analysis will of course suffer in its length, but perhaps could be offered as a path for future research to attempt. Creating a toolset to generate a highly-usable version of the Library of Babel could be useful for later researchers. If option two is attempted, then we will attempt to perform this research ourselves. Either way, we feel our project provides an opportunity to develop interesting results to a complex and abstract question.