

Troubleshooting Text Analysis

[Click to schedule a meeting with the DITI Team](#)

Because text analysis often involves a number of complex operations, when something goes wrong, it can be difficult to figure out what exactly has gone wrong and how to address the error. Like all technical work, errors and unexpected results are part of the process of learning how to use a tool, and sometimes, these moments can serve to both teach you more about the methods you're using as well as the text you're hoping to analyze. While the specifics of troubleshooting text analysis may vary depending on the tool you are using, this handout will cover some general guidelines which are shared across tools.

Identifying The Error

When you are using a text analysis interface tool such as Voyant or Lexos—that is, a tool where much of the computation is happening behind the scenes—errors are typically triggered by either an error with the interface itself or an issue with the text you are attempting to analyze. Since the text analysis computation is not customizable and doesn't require you, the user, to change any critical settings, it isn't likely that anything **you** have done has produced an error in the code which underpins the functionality of the tool.

Interface Errors

Interface errors occur when there is an issue with the tool itself. These errors may occur when the server is down, when the code running behind the scenes is no longer up to date, or when an update contains a bug. To address this form of error, you should:

- Try to run the tool on a different corpus to see if the error persists—this will verify that it is the interface which is causing the error
- Monitor any social media or documentation maintained by the tool where they announce updates or scheduled downtime—typically, interface errors will be announced there
- Check if other users are also experiencing the same issues on social media platforms, such as user forums for the software itself, Stack Overflow, Bluesky, or Reddit

Developed by: Avery Blankenship

Questions? Contact us: nulab.info@gmail.com

- Be patient and come back to the tool at a later time

Ultimately, interface errors are produced by the tool itself and are not errors that can be addressed by individual users. If the tool has a report feature or another way to report bugs or errors, you should do so in order to alert the team maintaining the tool that there is an issue in order to expedite a fix. If you **really** need to use the tool because of a time sensitive issue, you can attempt to load an earlier version of the tool using the [Wayback Machine](#), which captures a virtual image of the tool's state at an earlier point in time. The tool may not run as smoothly, but this method should allow you to use it in some limited capacity.

Corpus Errors

Corpus errors are produced by the corpus of text that you are attempting to analyze. Unlike interface errors, these errors can be addressed by the user immediately and are unrelated to the tool or interface. If you encounter an error with your corpus, you should:

- Run the tool on a different corpus to verify that the issue is with your corpus and not with the tool itself
- Verify that your text files are in the correct file format and if not, convert them to a format accepted by the tool
- Try to open one of the text files and take a look at the text itself—sometimes, the issue may be the text within the files (for instance, sometimes when text is converted from a .docx file to a .txt file, the text can become jumbled)
- Verify that the filenames aren't an issue—no repeating names, files are in the correct location, etc.
- Verify that the files are split the way that you want them to be—the text is either in a single machine-readable file or is split across multiple files. Similarly, ensure that files are either in a single folder or are embedded within a series of folders as is appropriate
- Remove any headers in the text file which might be producing unexpected results—for example, academic library headers or metadata headers which contain bibliographic data about the text you are analyzing
- Check the actual size of the text you are attempting to analyze; if your text files are too large, try splitting the files into smaller files
- Ensure that the files contains only text and no images
- Check the tool's documentation to ensure that the tool doesn't have specific requirements for corpus formatting

In most cases, one or most of these bullet points will be the culprit of the error and after addressing the issue, the tool will run as expected.

Fine-tuning Results

Text analysis, like much technical work, often involves running a tool a couple of times in order to get the “final” results. You may run a tool on a corpus an initial time just to get a sense of what the corpus looks like or what it is you’re actually working with, and then following that initial run, choose to finetune the corpus in order to produce more useful results. Text analysis is very rarely a method that is applied to a corpus only once. If you don’t get the most useful or intriguing results on an initial run of a tool on your corpus, some things you may consider in order to finetune your results (depending on the tool you are using):

- Increase the number of texts in your corpus
- Limit your corpus to a specific subject or genre of text
- Explore a central theme or word in more detail throughout the corpus—in other words, direct your analysis more narrowly
- Remove the [stop words](#) from your corpus, or other words which are likely to be repeated so often that they aren’t useful (for example, perhaps removing “whale” from *Moby Dick*)
- Remove character names in a corpus of novels if you aren’t interested in characters or character interactions

In other words, on a second run of the tool, you should think about the ways in which you can modify or alter your corpus in order to explore a theme or idea that you find interesting—perhaps a theme or idea that you became interested in or noticed was becoming obscured during that initial run.