Exploratory Text Analysis and Visualization of Text Data for Research with Open Source Tools

Scott Bailey

September 21, 2020

Contents

1 Presenter

Scott Bailey
Digital Research and Scholarship Librarian

Copyright and Digital Scholarship Center
NC State University Libraries

2 Exploratory text analysis and visualization in the research context

Key take-aways:

1. Text corpora exist with a wide range of metadata, cleanliness, and quality.
2. Even when we have a good sense of method or approach, exploring our data, whether quantitative or qualitative, helps us refine research questions, fine-tune methods, and scope our project to be successful.
3. Voyant and Taguette are free and open source tools that allow you to analyze and visualize texts, whether as single documents or whole corpora.

3 Voyant - Demo

3.1 Installing Voyant vs using the hosted web version

There is a hosted version of Voyant at https://voyant-tools.org/ that has two text corpora available. This is a great way to experiment with and learn Voyant, and can be used for your own corpora.

You can also download and host Voyant yourself, and I recommend this for working with larger corpora and in order to work offline. Since you can adjust the memory available to Voyant, there is an advantage in speed and the size of corpus you can analyze.

Download and install page: `https://digihum.mcgill.ca/voyant/resources/run-your-own/voyant-server/`

3.2 File types and corpus building

Voyant can handle single documents, either by uploading them or simply copying text into the Voyant home screen input box. It provides even more power, though, when you upload many documents as a corpus.

Supported file types:

- plain text: .txt

- HTML: .htm, .html
- XML: .xml
- Word: .doc, .docx
- RTF: .rtf
- PDF: .pdf

3.3 Exploring the interface

- Basic layout
- Getting help for each tool
- Settings for each tool
- Clicking on a document or word in one tool typically causes the other tools to update accordingly

3.4 Tools of note

- **Summary**: basic statistics on your document or your corpus, including most frequent terms and most distinctive terms per document (tf-idf)
- **Contexts**: keyword-in-context tool
- **Correlations**: Pearson's correlation for pairs of terms in the document/corpus, and significance for the correlation
- **Trends**: trends over documents or document chunks for terms according to relative frequency or raw count
- **Topics** (Hidden by default): topic modeling by way of the latent dirichlet allocation (LDA) algorithm
- There are many tools - explore them all!

3.5 Exporting results

What formats you can use for export depends on each tools, but in general:

- Exporting visualizations: PNG, HTML, SVG
- Exporting data: HTML, TSV, JSON

4 Taguette - Demo

4.1 Installing Taguette vs using the hosted web version

Taguette also has a hosted version at https://app.taguette.org/. You can use the hosted for full projects, including collaborating with others on your qualitative analysis project.

You can also download and run Taguette on your own computer, or run your own server. On your own computer, you won't be able to use the collaboration feature, but can still carry out your analysis and export results.

Download and install page: `https://www.taguette.org/install.html`

4.2 File types

Supported file types:

- plain text: .txt
- HTML: .htm, .html
- Word: .doc, .docx
- RTF: .rtf
- PDF: .pdf
- Open Documents: .odt
- E-books: .epub, .mobi

4.3 Basic workflow

- Creating a project
- Adding documents
- Highlighting text and adding tags
    - Creating hierarchical tags (parent and child codes)

4.4 Exporting results

- Exporting collated tagged texts: PDF, HTML, DOCX, XSLX, CSV
- Exporting your codebook: PDF, HTML, DOCX, XSLX, CSV, XML

5 Putting it together: between distant and close reading

Key take-aways:

1. Distant reading allows us to analyze and draw conclusions from collections of texts at large scale.
2. Close reading helps us develop a nuanced understanding of texts.
3. Research is iterative, and moving between distant and close reading allows us to develop interpretations and conclusions that draw on the whole of a corpus without losing the nuance.

6 Resources

- Voyant Docs: `https://digihum.mcgill.ca/voyant/`
- Voyant Tools Docs: `http://docs.voyant-tools.org/tools/`
- UC Santa Cruz Getting Started with Voyant: `https://guides.library.ucsc.edu/DSCguides/Voyant`
- Taguette Getting Started Guide: `https://www.taguette.org/getting-started.html`
- Illinois Library Coding with Taguette Guide: `https://guides.library.illinois.edu/qualitative/taguette`