Distant Reader Workshop Hands-On Activities

This is a small set of hands-on activities presented for the Keystone Digital Humanities 2021 annual meeting. The intent of the activities is to familiarize participants with the use and creation of Distant Reader study carrels.

Introduction

The Distant Reader is a tool for reading. Given an almost arbitrary amount of unstructured data (text), the Reader creates a corpus, applies text mining against the corpus, and returns a structured data set amenable to analysis ("reading") by students, researchers, scholars, and computers.

The data sets created by the Reader are called "study carrels". They contain a cache of the original input, plain text versions of the same, many different tab-delimited files enumerating textual features, a relational database file, and a number of narrative reports summarizing the whole. Given this set of information, it is easy to answer all sorts of questions that would have previously been very time consuming to address. Many of these questions are akin to newspaper reporter questions: who, what, when, where, how, and how many.

Using more sophisticated techniques, the Reader can help you elucidate on a corpus's aboutness, plot themes over authors and time, create maps, create timelines, or even answer sublime questions such as, "What are some definitions of love, and how did the writings of St. Augustine and Jean-Jacques Rousseau compare to those definitions?"

The Distant Reader and its library of study carrels are located at:

- https://distantreader.org
- https://distantreader.org/catalog

Activity #1: Compare & contrast two study carrels

These tasks introduce you to the nature of study carrels:

1. From the library, identify two study carrels of interest, and call them Carrel A and Carrel B. Don't think too hard about your selections.
2. Read Carrel A, and answer the following three questions: 1) how many items are in the carrel, 2) if you were to describe the content of the carrel in one sentence, then what might that sentence be, and 3) what are some of the carrel's bigrams that you find interesting and why.
3. Read Carrel B, and answer the same three questions.
4. Answer the question, "How are Carrels A and B similar and different?"

Activity #2: Become familiar with the content of a study carrel

These tasks stress the structured and consistent nature of study carrels:

1. Download and uncompress both Carrel A and Carrel B.
2. Count the number of items (files and directories) at the root of Carrel A. Count the number of items (files and directories) at the root of Carrel B. Answer the question, "What is the difference between the

two counts?". What can you infer from the answer?
3. Open any of the items in the directory/folder named "cache", and all of the files there ought to be exact duplicates of the original inputs, even if they are HTML documents. In this way, the Reader implements aspects of preservation. A la LOCKSS, "Lots of copies keep stuff safe."
4. From the cache directory, identify an item of interest; pick any document-like file, and don't think too hard about your selection.
5. Given the name of the file from the previous step, open the file with the similar name but located in the folder/directory named "txt", and you ought to see a plain text version of the original file. The Reader uses these plain text files as input for its text mining processes.
6. Given the name of the file from the previous step, use your favorite spreadsheet program to open the similarly named file but located in the folder/directory named "pos". All files in the pos directory are tab-delimited files, and they can be opened in your spreadsheet program. I promise. Once opened, you ought to see a list of each and every token ("word") found in the original document as well as the tokens' lemma and part-of-speech values. Given this type of information, what sorts of questions do you think you can answer?
7. Open the file named "MANIFEST.htm" found at the root of the study carrel, and once opened you will see an enumeration and description of all the folders/files in any given carrel. What types of files exist in a carrel, and what sorts of questions can you address if given such files?

Activity #3: Create study carrels

Anybody can create study carrels, there are many ways to do so, and here are two:

1. Go to https://distantreader.org/create/url2carrel, and you may need to go through ORCID authentication along the way.
2. Give your carrel a one-word name.
3. Enter a URL of your choosing. Your home page, your institutional home page, or the home page of a Wikipedia article are good candidates.
4. Click the Create button, and the Reader will begin to do its work.
5. Create an empty folder/directory on your computer.
6. Identify three or four PDF files on your computer, and copy them to the newly created directory. Compress (zip) the directory.
7. Go to https://distantreader.org/create/zip2carrel, and you may need to go through ORCID authentication along the way.
8. Give your carrel a different one-word name.
9. Select the .zip file you just created.
10. Click the Create button, and the Reader will begin to do its work.
11. Wait patiently, and along the way the Reader will inform you of its progress. Depending on many factors, your carrels will be completed in as little as two minutes or as long as an hour.
12. Finally, repeat Activities #1 and #2 with your newly created study carrels.

Extra credit activities

The following activities outline how to use a number of cross-platform desktop/GUI applications to read study carrels:

- Print any document found in the cache directory and use the traditional reading process to... read it. Consider using an active reading process by annotating passages with your pen or pencil.
- Download Wordle from the Wayback Machine, a fine visualization tool. Open any document found in

the txt directory, and copy all of its content to the clipboard. Open Wordle, paste in the text, and create a tag cloud.
- Download AntConc, a cross-platform concordance application. Use AntConc to open one more more files found in the txt directory, and then use AntConc to find snippets of text containing the bigrams identified in Activity #1. To increase precision, configure AntConc to use the stopword list found in any carrel at etc/stopwords.txt.
- Download OpenRefine, a robust data cleaning and analysis program. Use OpenRefine to open one or more of the files in the folder/directory named "ent". (These files enumerate named-entities found in your carrel.) Use OpenRefine to first clean the entities, and then use it to count & tabulate things like the people, places, and organizations identified in the carrel. Repeat this process for any of the files found in the directories named "adr", "pos", "wrd", or "urls".

Extra extra credit activities

As sets of structured data, the content of study carrels can be computed against. In other words, programs can be written in Python, R, Java, Bash, etc. which open up study carrel files, manipulate the content in ways of your own design, and output knowledge. For example, you could open up the named entity files, select the entities of type PERSON, look up those people in Wikidata, extract their birthdates and death dates, and finally create a timeline illustrating who was mentioned in a carrel and when they lived. The same thing could be done for entities of type GRE (place), and a map could be output. A fledgling set of Jupyter Notebooks and command-line tools have been created just for these sorts of purposes, and you can find them on GitHub:

- https://github.com/ericleasemorgan/reader-notebooks
- https://github.com/ericleasemorgan/reader-toolbox-classic/
- https://github.com/ericleasemorgan/reader-toolbox

Every study carrel includes an SQLite relational database file (etc/reader.db). The database file includes all the information from all tab-delimited files (named-entities, parts-of-speech, keywords, bibliographics, etc.). Given this database, a person can either query the database from the command-line, write a program to do so, or use GUI tools like DB Browser for SQLite or Datasette. The result of such queries can be elaborate if-then statement such as "Find all keywords from documents dated less than Y" or "Find all documents, and output them in a given citation style." Take a gander at the SQL file named "etc/queries.sql" to learn how the database is structured. It will give you a head start.

Summary

Given an almost arbitrary set of unstructured data (text), the Distant Reader outputs sets of structured data known as "study carrels". The content of study carrels can be consumed using the traditional reading process, through the use of any number of desktop/GUI applications, or programmatically. This document outlined each of these techniques.

Embrace information overload. Use the Distant Reader.

--
Eric Lease Morgan <emorgan@nd.edu>
July 10, 2021