homer-works-gutenberg

Eric Lease Morgan

Publication Date

19-12-2023

Distant Reader "study carrels": A manifest

The results of the Distant Reader process is the creation of a "study carrel" -- a set of structured data files intended to help you to further "read" your corpus. This document enumerates & outlines the contents of a study carrel. A future blog posting will describe ways to use & understand the files outlined here. Therefore, the text below is merely a manifest of a typical Distant Reader study carrel.

The Distant Reader takes an arbitrary amount of unstructured data (text) as input, and it outputs sets of structured data files for analysis -- reading. Given a corpus of any size, the Distant Reader will analyze the corpus, and it will output a myriad of reports enabling you to use & understand the corpus. The Distant Reader is intended to supplement the traditional reading process. Given a question of a rather quantitative nature, a Distant Reader study carrel may very well contain a plausible answer.

The results of downloading and uncompressing the Distant Reader study carrel is a directory/folder containing a standard set of files and subdirectories. Each of these files and subdirectories are listed and described below:

- A1426341535 - This, or a very similarly named file, is an administrative file, a unique identifier created by the system (Airivata) which processed the study carrel. [1] In the future, this file may not be included. On the other hand, since the file's name is a unique identifier, then it could be exploited by a developer.

- adr - This subdirectory contains a set of tab-delimited files. Each file contains a set of email addresses extracted from the documents in your corpus. While the files' names end in .adr, they are plain text files that can be imported into for favorite spreadsheet, database, or analysis application. The files have two columns: 1) id, and 2) address. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files can humorously answer the question "Who are you gonna call?"

- bib - This subdirectory contains a set of tab-delimited files. Each file contains a set of rudimentary bibliographic information from a given document in your corpus. While the files' names end in .bib, they are plain text files that can be imported into for favorite spreadsheet, database, or analysis application. The files have thirteen columns: 1) id, 2) author, 3) title, 4) date, 5) page 6), extension, 7) mime, 8) words, 9) sentences, 10) flesch, 11) summary, 12) cache, and 13) txt. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files help answer the question "What items are in my corpus, and how can they be described?"

- cache - This subdirectory contains original copies of the files you intended for analysis. It is populated by harvesting content from URLs or were supplied in the zip file you uploaded to the Reader. Each file is named with a unique and somewhat meaningful name and an extension. These files are intended for reading on your computer, or better yet, printed and then read in the more traditional manner.

- css - This subdirectory contains a set of cascading stylesheets used by

the HTML files in the carrel. If you really desired, one could edit these files in order to change the appearance of the carrel.

- input.zip - This file, or something named very similarly, is the file originally used to create your study carrel. It has already served its intended purpose, but it is retained for reasons of provenance.

- ent - This subdirectory contains a set of tab-delimited files, and each file contains a set of named entities from a given document in your corpus. While the files' names end in .ent, they are plain text files that can be imported into for favorite spreadsheet, database, or analysis application. The files have five columns: 1) id, 2) sid, 3) eid, 4) entity, and 5) type. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files help answer questions regarding who, what, when, where, how, and how many.

- etc - This subdirectory contains a set of ancillary files, and each are described below:
    - model-data.txt - the data file used by topic-model.htm, and it is essentially an enhanced version of reader.txt

    - queries.sql - a set of SQL queries used to generate report.txt, and this file is an excellent introduction to the use of reader.db

    - reader.db - an SQLite database file, and it is essentially the amalgamation of the contents of the adr, bib, ent, pos, urls, and wrd directories; the intelligent use of this file can be used to answer just about any question answerable by the carrel

    - reader.sql - a set SQL commands denoting the structure of reader.db

    - reader.txt - the concatenation of all files in the txt directory; a plain text version of the whole of the corpus is often used for other purposes and it is provided here as a convienence

    - report.txt - the result of applying queries.sql to reader.db; this file has the exact same content as standard-output.txt

    - stopwords.txt - a list of function words (i.e. "a", "an", "the", etc.) used through the creation of the study carrel

- figures - This subdirectory contains a set of image files used by the carrel's HTML files:
    - adjectives.png - a word cloud illustrating the most frequent adjectives in the corpus

    - adverbs.png - a word cloud illustrating the most frequent adverbs in the corpus

    - bigrams.png - a word cloud illustrating the most frequent bigrams (two-word phrases) in the corpus

    - flesch-boxplot.png - a box plot illustrating the average, quartile, and outlier readability scores of the items in the corpus

    - flesch-histogram.png - a histogram illustrating the distribution of readability scores of the items in the corpus

    - keywords.png - a word cloud illustrating the most frequent keywords (statistically significant unigrams) in the corpus

    - nouns.png - a word cloud illustrating the most frequent nouns in the corpus

    - pronouns.png - a word cloud illustrating the most frequent pronouns in the corpus
    - proper-nouns.png - a word cloud illustrating the most frequent proper nouns in the corpus
    - sizes-boxplot.png - a box plot illustrating the average, quartile, and outlier sizes of the items (measured in unigrams) in the corpus
    - sizes-histogram.png - a histogram illustrating the distribution of sizes of the items (measured in unigrams) in the corpus
    - topics.png - a pie chart illustrating how the corpus is subdivided if topic modeling were applied to the corpus, and the desired number of topics (latent themes) equals five
    - unigrams.png - a word cloud illustrating the most frequent unigrams (individual words) in the corpus
    - verbs.png - a word cloud illustrating the most frequent verbs in the corpus
- htm - This subdirectory contains a set of interactive HTML files linked from the file named index.htm. The functionality of each file is outlined below:
    - adjective-noun.htm - search, sort, and browse adjective/noun combinations by adjective, noun, or frequency
    - adjectives.htm - search, sort, and browse adjectives and/or their frequency
    - adverbs.htm - search, sort, and browse adverbs and/or their frequency
    - bigrams.htm - search, sort, and browse bigrams (two-word phrases) and/or their frequency
    - entities.htm - search, sort, and browse named-entities, their type, and/or their frequency
    - keywords.htm - search, sort, and browse keywords (statistically significant unigrams) and/or their frequency
    - noun-verb.htm - search, sort, and browse noun/verb combinations by noun, verb, or frequency
    - nouns.htm - search, sort, and browse nouns and/or their frequency
    - pronouns.htm - search, sort, and browse pronouns and/or their frequency
    - proper-nouns.htm - search, sort, and browse proper nouns and/or their frequency
    - quadgrams.htm - search, sort, and browse quadgrams (four-word phrases) and/or their frequency
    - questions.htm - search, sort, and browse questions (sentences ending with a question mark) and from which items they were extracted
    - search.htm - a free text query interface based on the narrative summaries of each item in the corpus
    - topic-model.htm - a topic modeler; a tool used to enumerate as well

as compare & contrast latent themes in the corpus

  - trigrams.htm - search, sort, and browse trigrams (three-word phrases) and/or their frequency
  - unigrams.htm - search, sort, and browse unigrams (individual words) and/or their frequency
  - verbs.htm - search, sort, and browse verbs and/or their frequencies

- index.htm - This HTML file narratively reports on the content of your study carrel. It is the best place to begin once you have downloaded and unzipped the carrel.
- MANIFEST.htm - This file, and it is the third best place to begin once you have downloaded and unzipped a carrel.
- job_1819387465.slurm - This file, or a very similarly named file, is the batch file used to initially create your study carrel. In the future, this file may be removed from the study carrel all together because it serves only an administrative purpose.
- js - This subdirectory includes a set of Javascript libraries supporting the functionality of index.htm as well as the HTML files in the htm directory. Because these files are here your computer does not need to be connected to the Internet in order to effectively read your carrel. Study carrels are designed to be stand-alone file systems usable for years to come.
- LICENSE - This is the license file; each study carrel is distributed under a GNU Public License.
- pos - This subdirectory contains a set of tab-delimited files, and each file contains a set of part-of-speech files from a given document in your corpus. While the files' names end in .pos, they are plain text files that can be imported into for favorite spreadsheet, database, or analysis application. The files have six columns: 1) id, 2) sid, 3) tid, 4) token, 5) lemma, and 6) pos. The definitions of these columns are described in another blog posting. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files help answer question regarding who, what, how, how many, and actions as well as grammer and style.
- README - This file contains the *very briefest* of introductions to the carrel.
- standard-error.txt - As each study carrel is being created, error and status messages are output to this file. It is a log file. If the creation of your study carrel fails, then this is a good place to look for clues on what went wrong. Send me this file if you are stymied.
- standard-output.txt - After your study carrel as been created and distilled into a database, sets of queries are applied against the database. This file is the second best place to begin once you have downloaded and unzipped a carrel.
- tsv - Except for one (questions.tsv), this subdirectory contains a set of frequency tables in the form of tab-delimited text files. The exception is a tab-delimited text file too, but it is just not a frequency file. All of these files can be imported into for favorite spreadsheet, database, or analysis application. Possible uses for these files are destined to be outlined in future postings, but in short, perusal of these files will

help you answer questions regarding your corpus's "aboutness" as well as who, what, when, where, how, how many, and why questions. The structure of each file is listed below:

  - adjective-noun.tsv - three columns: 1) adjective, 2) noun, and 3) frequency where frequency denotes the number of times the given adjective appears immediately before the given noun in the corpus
  - adjectives.tsv - two columns: 1) adjective, and 2) frequency
  - adverbs.tsv - two columns: 1) adverb, and 2) frequency
  - bigrams.tsv - two columns: 1) bigram (two-word phrase), and 2) frequency
  - entities.tsv - three columns: 1) entity, 2) type, and 3) frequency
  - keywords.tsv - two columns: 1) keyword (statistically significant unigram), and 2) frequency
  - noun-verb.tsv - three columns: 1) noun, 2) verb, and 3) a frequency where frequency denotes the number of times the given noun appears immediately before the given verb in the entire corpus
  - nouns.tsv - two columns: 1) noun, and 2) frequency
  - pronouns.tsv - two columns: 1) pronoun, and 2) frequency
  - proper-nouns.tsv - two columns: 1) proper, and 2) frequency
  - quadgrams.tsv - two columns: 1) quadgram (four-word phrase), and 2) frequency
  - questions.tsv - two columns: 1) identifier, and 2) question where each question is a "sentence" ending in a question mark
  - trigrams.tsv - two columns: 1) trigram (three-word phrase), and 2) frequency
  - unigrams.tsv - two columns: 1) unigram (individual word), and 2) frequency
  - verbs.tsv - two columns: 1) verb, and 2) frequency

- txt - This subdirectory contains plain text versions of the files stored in the cache directory. A plain text version of each & every item in the cache directory ought to exist in this directory. The contents of this directory is what was used to do the Reader's analysis. The contents of this directory are excellent candidates for further analysis with tools such as concordances, indexers, or topic modelers.

- urls - This subdirectory contains a set of tab-delimited files, and each file contains a set of URLs from a given document in your corpus. While the files' names end in .url, they are plain text files that can be imported into for favorite spreadsheet, database, or analysis application. The files have three columns: 1) id, 2) domain, and 3) url. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files help answer questions regarding document provenance and relationships as well as addressing the perenial issue of "finding more like this one".

- wrd - This subdirectory contains a set of tab-delimited files, and each file contains a set of computed keywords from a given document in your corpus. While the files' names end in .wrd, they are plain text files that

can be imported into for favorite spreadsheet, database, or analysis application. The files have two columns: 1) id, and 2 keyword. The definitions of these columns and possible uses of these files are described elsewhere, but in short, these files help answer questions such as "What is this document about?"

Links

[1] Airivata - https://airavata.apache.org

Eric Lease Morgan <emorgan@nd.edu>
December 26, 2019