Digital Humanities in Practice

WEEK 5a & 5b: Corpus Building/Preparing Texts for Analysis

Text Cleaning

Much of the text cleaning work you do in this class will be experimental - you may toggle cleaning options on and off to determine what effect they have on your content set. *You will not end up with perfectly cleaned documents!* In order to do this, it would be necessary to intervene with manual cleanup which is not an option currently in the DSL. It also takes a very long time!

The following reading will provide a framework for understanding WHY we clean data:

Katie Rawson, Trevor Muñoz, "Against Cleaning", http://curatingmenus.org/articles/against-cleaning/, July 6 2016

Miriam Posner, “Humanities Data: A Necessary Contradiction”, https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/, June 25 2015

WORDS TO NUMBERS: TEXT PREPROCESSING CHOICES

The following excerpt lists seven of the textual features that you may want to consider removing from your text in order to conduct a meaningful analysis. Essentially, we are working with a 'bag of words' model which is often mined quantitatively to extract patterns and frequencies. Many of these choices can be made by checking the appropriate box on the DSL cleaning page, but this list will give you some understanding of why you might want select one or more cleaning options.

Punctuation: The first choice a researcher must make when deciding how to preprocess a corpus is what classes of characters and markup to consider as valid text. The most inclusive approach is simply to choose to preprocess all text, including numbers, any markup (html) or tags, punctuation, special characters ($, %, &, etc), and extra white-space characters. These non-letter characters and markup may be important in some analyses (e.g. hashtags that occur in Twitter data), but are considered uninformative in many applications. It is therefore standard practice to

remove them. The most common of these character classes to remove is punctuation. The decision of whether to include or remove punctuation is the first preprocessing choice we consider.

Numbers: While punctuation is often considered uninformative, there are certain domains where numbers may carry important information. For example, references to particular sections in the U.S. Code ("Section 423", etc.) in a corpus of Congressional bills may be substantively meaningful regarding the content legislation. However, there are other applications where the inclusion of numbers may be less informative.

Lowercasing: Another preprocessing step taken in most applications is the lowercasing of all letters in all words. The rationale for doing so is that that whether or not the first letter of a word is uppercase (such as when that words starts a sentence) most often does not affect its meaning. For example, "Elephant" and "elephant" both refer to the same creature, so it would seem odd to count them as two separate word types for the sake of corpus analysis. However, there are some instances where a word with the same spelling may have two different meanings that are distinguished via capitalization, such as "rose" (the flower), and "Rose" the proper name.

Stemming: The next choice a researcher is faced with in a standard text preprocessing pipeline is whether or not to stem words. Stemming refers to the process of reducing a word to its most basic form. For example the words "party", "partying", and "parties" all share a common stem "parti". Stemming is often employed as a vocabulary reduction technique, as it combines different forms of a word together. However, stemming can sometimes combine together words with substantively different meanings ("college students partying", and "political parties"), which might be misleading in practice.

Stopword Removal: After tokenizing the text, the researcher is left with a vector of mostly meaningful tokens representing each document. However, some words, often referred to as "stop words", are unlikely to convey much information. These consist of function words such as "the", "it", "and", and "she", and may also include some domain-specific examples such as "congress" in a corpus of U.S. legislative texts. There is no single gold-standard list of English stopwords, but most lists range between 100 and 1,000 terms. Most text analysis software packages make use of a default stopword list which the software authors have attempted to construct to provide "good performance" in most cases. There are an infinite number of potential stopword lists, and for this class we are using the Glasgow list.

n-gram Inclusion: While it is most common to treat individual words as the unit of analysis, some words have a highly ambiguous meaning when taken out of context. For example the word

“national” has substantially different interpretations when used in the multi-word expressions: “national defense”, and “national debt”. This has lead to a common practice of including n-grams from documents where an n-gram is a contiguous sequence of tokens of length n. For example, the multi-word expression “a common practice” from the previous sentence would be referred to as a 3-gram or tri-gram (assuming stopwords were not removed). Previous research has tended to use 1,2,and 3-grams combined, because this combination offers a reasonable compromise between catching longer multi-word expressions and keeping the vocabulary relatively smaller. After extracting all n-grams from a document, a number of approaches have been proposed to filter the resulting n-grams, but here we choose to focus only on the most basic case of considering all 1,2, and 3-grams together without any filtering. So, the decision of whether include 2 and 3-grams (along with unigrams, which are always included) is the sixth preprocessing choice we consider.

Infrequently Used Terms: In addition to removing common stopwords, researchers often remove terms that appear very infrequently as part of corpus preprocessing. The rationale for this choice is often two-fold; (1) theoretically, if the researcher is interested in patterns of term usage across documents, very infrequently used terms will not contribute much information about document similarity. And (2) practically, this choice to discard infrequently used terms may greatly reduce the size of the vocabulary, which can dramatically speed up many corpus analysis tasks. A commonly used rule of thumb is to discard terms that appear in less than 0.5-1% of documents, however, there has been no systematic study of the effects this preprocessing choice has on downstream analyses. The decision of whether include or remove terms that appear in less than 1% of documents is the seventh and final preprocessing choice we consider.

List excerpted and adapted from Denny, Matthew and Spirling, Arthur, Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It (September 27, 2017), p6-9. Available at SSRN: https://ssrn.com/abstract=2849145 or http://dx.doi.org/10.2139/ssrn.2849145

TEXT CLEANING IN THE DIGITAL SCHOLAR LAB

Here's the 'Cleaning' DSL video: https://youtu.be/nDNyM6KPxcA

- In the Digital Scholar Lab, continue to search and build Content Set(s) related to your chosen research topic.
- In the Doc Explorer view, compare the side-by-side original image with its OCR text output carefully. Can you identify any recurrent errors in the text?

- Begin the process of creating a custom Cleaning Configuration to apply to your content set.
- Test your cleaning configuration by following the 'test configuration' process, downloading 10 cleaned and 10 uncleaned documents NOTE: THIS IS SLOW, ITERATIVE WORK!
- Repeat the process of tweaking your configuration, and re-testing.

From the DSL 'Help' Pages

Cleaning & Configuration

One of the most important elements of text analysis is making sure that your texts are formatted in a way that suits the kind of analysis you want to carry out. The Clean feature of the Digital Scholar Lab lets you edit all the Documents within a Content Set. It's designed to fit into existing Analysis Tools, as well as the download process for a Content Set. While we can edit or alter Documents on the fly for particular tools, cleaning is a critical part of the preparation for any text analysis. The Digital Scholar Lab breaks it out as a separate feature, so you can ensure that the Documents in different Content Sets are prepared in precisely the same way, and that you, as a scholar, can decide how they're altered, and make adjustments according to your research needs. It will be a one-stop shop for tinkering with the texts you want to Analyze before sending them off to an Analysis Tool Job.

Configurations & Replication / Method

The Clean feature is designed as part of the broad commitment to transparency and method within the Digital Scholar Lab. We want you to be able to replicate or reproduce your results with the same Content Set, or compare similar methods across different Content Sets. Clean lets you build a Configuration you can reuse, alleviating the need to remember 'what you did', as much as allowing you to return to your analysis easily after being away from the Digital Scholar Lab for a period of time. In short, a Configuration creates a kind of standardized method for preparing documents that you can send for Analysis, and lets you use that standardized preparation - like a cookie cutter - for any of your Content Sets, combined with any Analysis Tool.

Default & Custom Configurations

Clean allows you to create Configurations which you can reuse or associate with a specific Analysis Job for an Analysis Tool. Default Configurations can be used directly, or they can act as a template or starting point for the creation of new Custom Configurations. As new Tools come online we'll provide you with a series of Default Configurations best suited to specific Tools,

which you can tweak and alter as you see fit. You can save the Configurations at any time - just provide a new name, and description, which will help you pick the correct Configuration from a dropdown list in a Tool configuration area.

Select the text cleaning configurations

Each Configuration consists of a series of correct, removal and replacement or substitute options, alongside a possible stop word list. In theory, you can have a Configuration that's empty - it won't do anything, but the algorithm would still use it before running an Analysis Job.

Corrections

- The only correction currently available is case correction, altering all text to lower case. This is useful in contexts where an Algorithm or Tool might be case sensitive and have no internal options to alter cases.

Removal

- Remove all number characters
- Remove all special characters. Users can set specific special characters to remove, such as currency symbols, slashes, underscores etc. Remove all punctuation. Users can set specific punctuation to remove

Replacement

- Reduce multiple spaces to one space (ex: "hello there" becomes "hello there")
- Replace _____ with _____ allows users to define what kinds of replacements they'd like to make on the fly. This function is useful for controlling orthography or spelling variants, e.g. all instances of 'colour' can be altered to 'color', to make sure that the Analysis Tools treat them as the same token, not distinct words.

Stop words

You can also decide to use a stop word list as part of your configuration. The default stop word list contains English words; however, you can edit this list as you see fit. Each stop word should be listed on a separate line. If you'd like, you can cut and paste an entirely new stop word list here.

Configurations and Tools

You can use any configuration on any Tool in the DS Lab environment. That said, different selections will impact certain tools in specific kinds of ways. For instance, MALLET - the software powering the Topic Modeling Tool - is case sensitive. If you decide to make everything lower case, it won't distinguish between Smith (perhaps someone's last name), and smith (an occupation, like a blacksmith). MALLET also fails handles possessive apostrophes in slight awkward manner, turning them into their own words - you can add 's to the Stop Word list to prevent this from happening. In some Sentiment Analysis Tools, punctuation, specifically periods, matter because they carve up Documents by Sentence. In the current version of the Sentiment Analysis Tool, this doesn't matter, but later versions will allow you to calculate Sentiment Scores in different ways for a Document. Removing all special characters will also remove currency symbols like $, £, ¥, or €, which might be important for tracking amounts. The Ngram tool, and many others, Tokenize, or cut up, Documents using whitespaces. It's prudent to replace all tabs or other characters (that might have slipped by our OCR processing) with single spaces to make sure that Documents are Tokenized appropriately. Because of this, we've pre-selected the whitespace options in our Default Configuration; you can change these, if you'd like.

In the end, there are no 'incorrect' Configuration options. But it is important to note how certain choices will affect or shape the results of an Analysis Job. Often the best method is a combination of reading up on what kinds of analyses you'd like to carry out, understanding how certain preparations can affect them, and testing them out. If you find something isn't quite right, the great thing is that you can change your Configuration, and run the Job again!

Name and save the cleaning configurations

When you first open up the Clean feature you'll see the 'Default Configuration', which contains an English stop word list, and a couple of configurations checked. You can alter this Default and save it as something different, simply provide a new Name, and if you think it's useful, a description. Each time you view and alter a Configuration, you can either save it, rewriting over the existing Configuration, or you can save it as something new, using save as. This allows you to create any number of Configurations for your work! Remember, though, to name them so you can remember the settings you've chosen! Using a kind of shorthand in the name helps, e.g. No Punctuation - Lower case - English for a Configuration which removes all punctuation, transforms all text to lower case, and uses the English stop word list.

Review the output of the cleaned content set

Most researchers will want to see what a Cleaning Configuration does to their texts before using it on a large Analysis Job. This is understandable - often it helps to look at edits to appreciate how

they might affect a larger computational task even though we have an expectation of what they should do, it's nice to know what the output might look like.

Sample

To check out your Configuration, just click on Test Configuration in the top Control Bar, and select the Content Set you'd like to test it on. This will submit the Cleaning Job and return a sample of 10 documents (original and cleaned texts; 20 documents total) from the Content Set for you to review. If you like what you see, you're good to go! If you don't like the results, you can then make changes accordingly, and rerun the test.

Content set up to 5,000 documents

At the moment Downloads are limited to 5,000 Documents per Content Sets. For Content Sets over 5,000 Documents, a Download will contain a randomized 5,000 Documents from your Content Set. You may apply a cleaning configuration at point of download to ensure you receive a cleaned version of your selected texts up to 5,000 documents.

Pass it through an analysis tool to generate a visualization

The last step for Clean is putting your Configuration into action. When you select an Analysis Tool, click on Configure, and you'll find your Configurations listed in the drop-down box at the top of the Configuration panel. Select the one you'd like to use for this tool, and then Run your Analysis Job as usual.

Cleaning Practice outside the Digital Scholar Lab

Option 1: Researchers have the option of downloading up to 5000 documents of OCR text from the DSL to clean or analyze outside the platform. While this allows for a little more flexibility in cleaning individual documents at a granular level, at the moment it's not possible to re-upload your cleaned documents in the DSL. So if you download your .txt files, chances are that you plan to analyze them outside of the DSL. However, for the purposes of this Module, it will be a useful exercise to explore a different cleaning options outside the DSL which are commonly used by researchers in digital humanities. There are three options listed below, and you're welcome to explore each of them.

Start by downloading your uncleaned content set (assuming it's less than 5000 documents) to your computer. You can do this from the 'My Content Sets' page in the Digital Scholar Lab.

Option 2: Use the data/primary source diaries and letters in the Data folder.

Lexos

Lexos is an open source, web-based tool which enables the researcher to upload, clean and analyze text material. It was developed and is maintained by Wheaton College.

From the Wheaton College Lexomics Department Website:

"Lexos is a web-based tool to help you explore your favorite corpus of digitized texts. Our primary motivation is to help you find the explorer spirit as you apply computational and statistical probes to your favorite collection of texts. Lexos provides a workflow of effective practices so you are mindful of the many decisions made in your experimental methods."

Full details are available here, including options to install on local machine. For the purposes of this class, I recommend their server instance:

http://lexos.wheatoncollege.edu/upload

The video tutorials are comprehensive and well worth watching. The full playlist is available here.

Summary: In this activity we will explore the impact of text cleaning on our data sets using the Lexos tool.

Instructions:

1. Open up your selected primary source documents from the Data folder and/or use OCR text downloaded from the Digital Scholar Lab.
2. Review the documents and consider any issues with the text. Consider how you would address those issues to get the best result.
3. Open up the Lexos tool using this link: http://lexos.wheatoncollege.edu/upload
4. Upload your data set by clicking the "Browse" button and selecting the files in your folder. You can also drag and drop the individual files into the "Drag files here" container. Successful uploads will appear in the "Upload List."

5. Once your files have been uploaded, click the “Prepare” link in the upper right hand corner and click the “Scrub” option in the drop down list.
6. Take a moment to experiment with the options on the page.
    a. Use the Previews section and click the “Preview” button to review your changes and see the impact on the text.
    b. Here are steps you can follow to use the Stop/Keep Words option
        i. Download the English_StopWord_List.txt.file from your Group folder
        ii. Paste the words in the English_StopWord_List.txt file in the text box
        iii. Click the “Stop” option
    c. Navigate to the “Previews” section on the page and click “Preview” button to see the changes to your text files.
    d. Tweak the English_Stop_Words.txt file and repeat the steps above until you’re happy with your results.
    e. Once you feel the text is in a good state, click the “Apply” button.
7. Now navigate to “Visualize” and select the Word Cloud option.
    a. Examine the terms in the word cloud
8. Navigate to the Scrub page again and download your cleaned data set by clicking the “Download” button in the upper right hand corner of the “Preview” section

The next two methods - Regular Expressions and Open Refine are considered 'medium difficulty'. Don't let this put you off! The Programming Historian tutorials will walk you through the process in each case and give you some exposure to methodologies you have perhaps not tried before. Make sure you take notes as you work, both for your assignment and as general good practice so that if you make a mistake, you can backtrack and pick up at the point where things went wrong.

Regular Expressions

The OCR texts you download from the DSL will be .txt files, or plain text. Plain text documents contain no hidden extra code. Word processor documents (.doc files, for example) can only be opened with word processors. Plain text documents can be opened with all and any text editors. Recommended editors which are free, even if you do get occasional popup boxes asking if you want to purchase include Sublime, Atom.io , BBEdit and Notepad++

Tutorial: Laura Turner O'Hara, "Cleaning OCR'd text with Regular Expressions ," The Programming Historian 2 (2013).

Another tutorial to consider: https://www.regular-expressions.info/quickstart.html

Additional Resources:

Beth Seltzer, "Text Scrubbing Hacks: Cleaning Your OCRed Text", https://sites.temple.edu/tudsc/2014/08/12/text-scrubbing-hacks-cleaning-your-ocred-text/

'Understanding Regular Expressions' https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions

Open Refine

Find Open Refine here, along with a selection of tutorials. Per the developers, it is: a *free, open source, powerful tool for working with messy data*

While you may not have time to appreciate all its features this course, the following tutorial will give you a taster.

Tutorial: The Programming Historian 'Text Cleaning with Open Refine ' (Seth van Hooland, Ruben Verborgh, and Max De Wilde, 2013)