INTRODUCTION TO DIGITAL HUMANITIES

SESSION FIVE

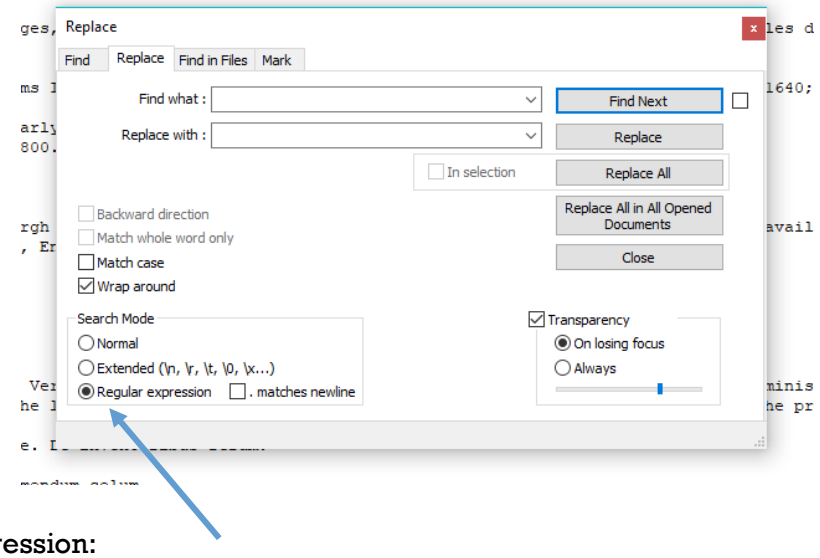
STATISTICAL ANALYSIS

INTRODUCTION

There are two activities this week. The first takes you through advanced find and replace to clean a data file using regular expressions. The second explores Voyant Tools.

ACTIVITY 1 – USE REGULAR EXPRESSIONS IN NOTEPAD++ (WINDOWS) OR TEXTMATE (APPLE) TO CLEAN DATA

Download the ESTC records text file **dataset-translations-estc-date-order.txt** from the resources for this session (these are all the records retrieved in session 2, downloaded in date order) and save it on a local drive. Open it in NotePad++ or TextMate and go to Search → Replace (NotePad++) or Edit → Find → Find (TextMate). If you don't have NotePad++ or TextMate installed on your computer, see the instructions at the end of this document.
Make sure the "Regular Expression" option is selected:



Search for this expression:

^\s*Record.*$

And replace it with nothing. Replace all.

@TAYOXFORD EMMA.HUBER@BODLEIAN.OX.AC.UK +44 1865 (2)78153

This expression means:

Find the start of a line (^ means start of line) followed by 0 or more (*) space characters (\s) followed by the word "Record". This is followed by 0 or more (*) characters (.) followed by the end of a line ($ means the end of a line).
We are replacing all that text with nothing at all. See what happens to your file!
Now try this:

> Find: ^\s*Title.*$
> Replace: nothing (an empty box)

We could keep going with (don't do this!):

^\s*Author.*$	^\s*Physical.*$	^\s*Added.*$
^\s*Uniform.*$	^\s*Surrogates.*$	^\s*Citation.*$
^\s*General.*$	^\s*Loc.*$	^\s*Collective.*$
^\s*Uncontrolled.*$	^\s*Copies.*$	^\s*Edition.*$
^\s*Variant.*$	^\s*Publisher.*$	^\s*Genre.*$

However, we can combine all of these into one expression, by creating a set using ( ) and a vertical pipe | which means or:

> ^\s*(Record|ESTC|Title|Author|Uniform|General|Uncontrolled|Variant|Physical|Surrogates|Loc|Copies|Publisher|Added|Citation|Collective|Edition|Genre)+.*$

Copy and paste this expression into the search box, make sure there is nothing in the replace box, and replace all.

All of the information we are not interested in is now deleted from the file. Of the remaining text, we want to retain the *text* of the subject line, but not the *label* "subject", "person as subject" or "corporate subject". To do this we need to select the text we want to keep using ( ) to create a set. The set contains all the characters on the line which come after the word subject.

> ^\s*(Subject|Corporate subject|Person as subject)+\s*(.*)$
> Replace with \2 (NotePad++) or $2 (TextMate)

\2 (or $2) pastes the second set in the find expression into your replacement – in this case, everything that appears after the label. Your file should now only contain the subjects, plus the header and footer which can easily be deleted by hand. Save as dataset-translations-estc-SUBJECTS-ONLY.txt (there is a ready-made version of this in the resources). For more information on regular expressions try exploring

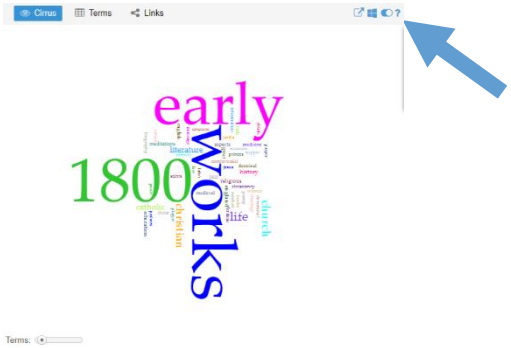
- https://regexone.com
- https://www.regular-expressions.info/reference.html
- https://regex101.com/



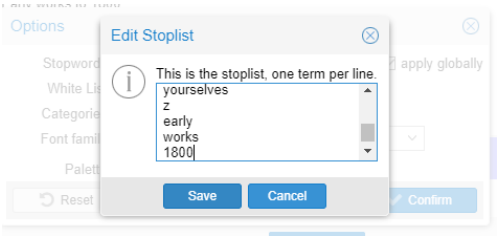
@TAYOXFORD **EMMA.HUBER@BODLEIAN.OX.AC.UK** **+44 1865 (2)78153**

ACTIVITY 2 – EXPLORE VOYANT-TOOLS

Go to https://voyant-tools.org/
Click on upload, and upload the text file which you created in activity one: **dataset-translations-estc-SUBJECTS-ONLY.txt**. There is also a ready-made version to download in the resources for this session.



The word cloud that you will see is heavily influenced by the phrase "early works to 1800" which is in nearly every subject, and tells us very little. Hover over the question mark on the top of the word cloud, and click on the switch button to define options for that tool. Next to "stopwords" click "edit" and add early, works and 1800. Click save.



See how the word cloud changes. This gives a better overview of the subjects of the texts in our corpus.

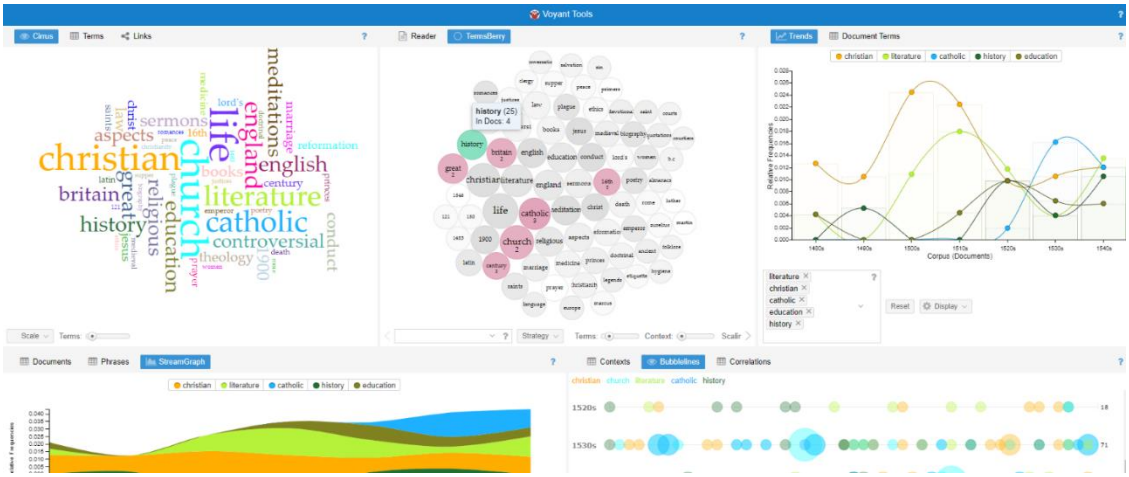
Because our document is in date order, we can see how the topics change over time:

@TAYOXFORD EMMA.HUBER@BODLEIAN.OX.AC.UK +44 1865 (2)78153

In the above example I have selected the visualisations "TermsBerry" which allows you to explore how different terms relate to each other; "Trends", which shows how usage of different terms changes over time (note how "Catholic" peaks, but then nearly disappears towards the end of the document); StreamGraph and Bubblelines, which are different visualisations of the same data.

Select terms using the dropdown menu for each tool, and select different tools by hovering over the windows icon which appears when you hover over the ? above each tool. Help for all the tools is available here: https://voyant-tools.org/docs/#!/guide/start

You may have already realised that relying on the document being in date order isn't a very accurate way of charting changing subjects over time. Try uploading the zip file "Dataset by decade – subject only.zip" to Voyant tools. This has all the same records, with the exception of those from 1550, but split into separate text files for each decade.



This has a very significant effect on our analysis – see how the term Catholic actually spikes towards the later decades. This demonstrates how careful we need to be with our methodology when using tools like this. We need to be aware of the limitations of our data, and of the tools we are using. Do explore the different tools with this dataset, and reflect on whether the analysis allows us to conclude anything, or suggests avenues for further research.

Other analysis tools you may wish to explore include:

- VOSviewer - http://www.vosviewer.com/
- AntConc - http://www.laurenceanthony.net/software/antconc/
- EPPI Reviewer - https://eppi.ioe.ac.uk/CMS/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4&

HOW TO INSTALL NOTEPAD++ OR TEXTMATE

NotePad++ is a very popular text editor, but it only works on Windows. Download it from https://notepad-plus-plus.org/downloads/ (click on the latest version, and then on the "Installer" for 32 or 64 bit).

On a Mac, try TextMate, which can also do regular expressions. You can download TextMate here: https://api.textmate.org/downloads/release?os=10.12

@TAYOXFORD EMMA.HUBER@BODLEIAN.OX.AC.UK +44 1865 (2)78153