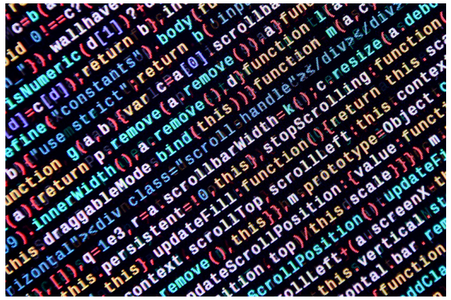
Machines of Knowledge - Maastricht University - November 2022

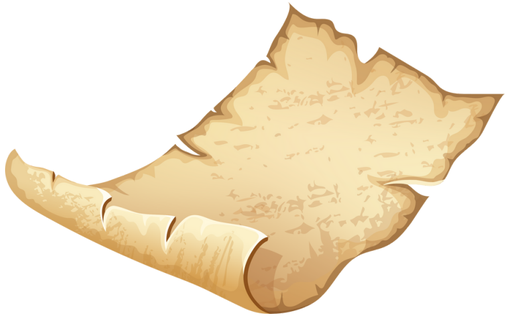
Text Analysis I: corpus building & tools

What we will do today...

- teaser for Costas's "distant reading" lecture
- corpus building as the basis of text analysis
- introduction to distant reading tools
- interpreting NLP visualisations

What is text analysis?





猴年大吉

Ways to engage with texts:
"close reading" &
"distant reading"
(phrase coined by
Franco Moretti)

What most humanities scholars do:
"middle distance reading"
(Marti Hearst et al.)

Handling terminology…

- **text mining/text analytics** = artificial intelligence (AI) technology to transform unstructured text in documents and databases into normalized, structured data suitable for automated analysis
- **distant reading** = (quantitative) reading method that relies on computer programs
- **natural language processing** (NLP) = field of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language
- **machine learning** (ML) = method of data analysis that automates analytical model building, making the computer "learn" from studying data and statistics

Aspects of distant reading:

- word frequencies
- narrative structures
- topic models (co-occurrences)
- sentiment analysis (speech tagging)
- stylometry (linguistic features, e.g. those specific to one author)
- named entity recognition (locate and classify people, places, or concepts)
- keywords in context

Corpus building as the basis of text analysis

What is a corpus (in NLP)?

"A corpus can be defined as a **collection of machine-readable authentic texts** (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety (McEnery et al. 2006: 5), though “representativeness” is a fluid concept (see Section 7.3)."

Xiao, R. (2010) Corpus creation. In N. Indurkhya & F. Damerau (eds) The Handbook of Natural Language Processing (2nd ed.), 147-165 . London: CRC Press.

Why is building a corpus challenging?

- "importance" of individual texts negotiable
- access to texts limited
- texts incomplete or lost
- human and computational ability to process multiple languages limited

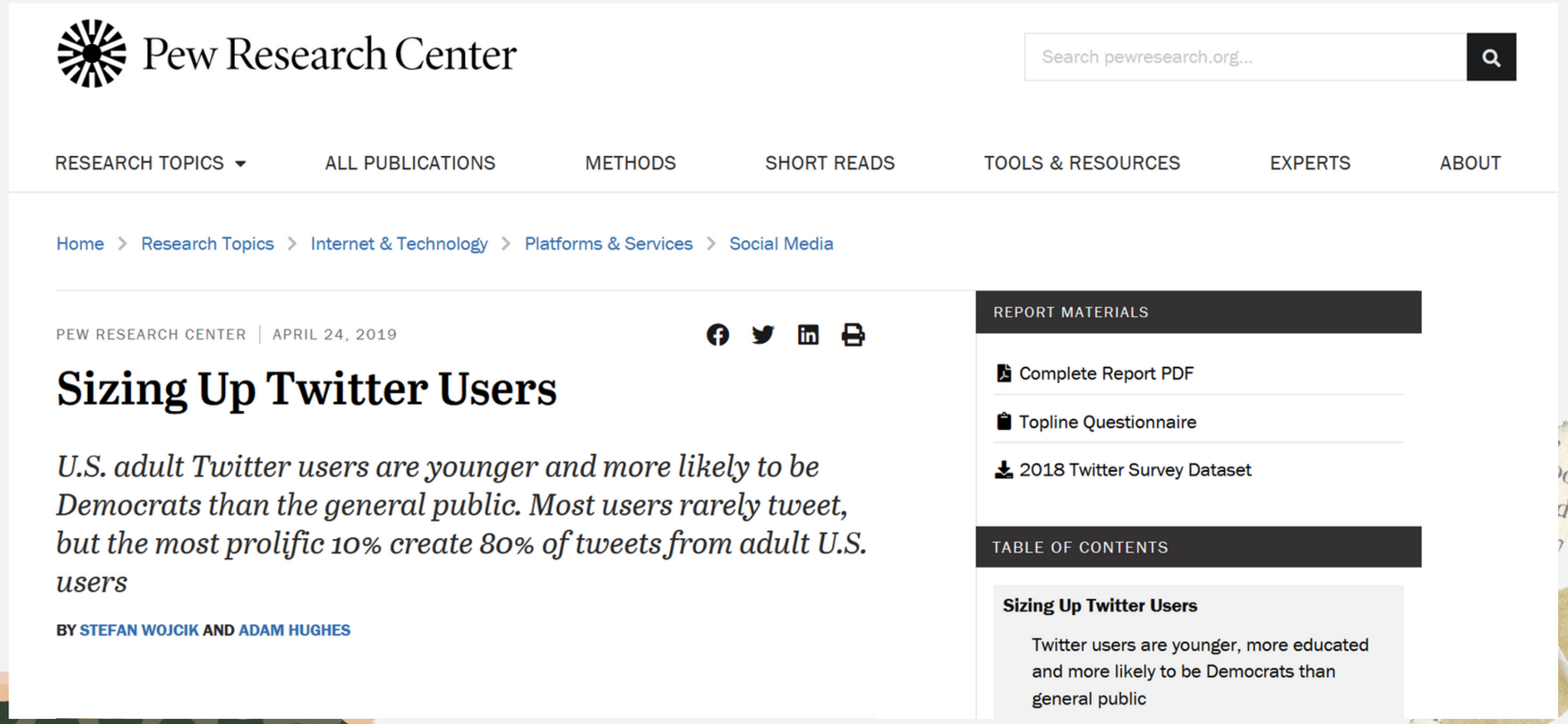
Where to find data for text analysis...

- HathiTrust (mostly older publications)
- Europeana (different European heritage collections)
- National archival services (e.g. LOC in the US)
- corpora provided by researchers / departments
- GITHUB (mainly processed data & transcriptions)
- social media (limited access)

Challenges of using social media for NLP:

- rate limits
- time restrictions
- ethical issues
- inherent biases

INHERENT DATA BIASES - *example:* TWEETS (FROM THE US)



Pew Research Center

Search pewresearch.org...

RESEARCH TOPICS ALL PUBLICATIONS METHODS SHORT READS TOOLS & RESOURCES EXPERTS ABOUT

Home > Research Topics > Internet & Technology > Platforms & Services > Social Media

PEW RESEARCH CENTER | APRIL 24, 2019

Sizing Up Twitter Users

U.S. adult Twitter users are younger and more likely to be Democrats than the general public. Most users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users

BY STEFAN WOJCIK AND ADAM HUGHES

REPORT MATERIALS

Complete Report PDF

Topline Questionnaire

2018 Twitter Survey Dataset

TABLE OF CONTENTS

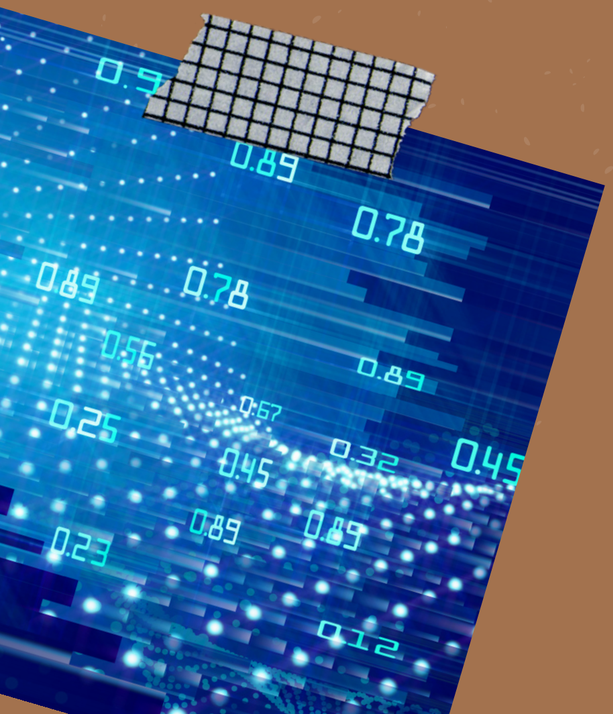
Sizing Up Twitter Users

Twitter users are younger, more educated and more likely to be Democrats than general public

Analysis by the Washington DC-based think tank "Pew Research Center" in April 2019:
https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Unreliable data - reliable research

While it is often impossible to find "complete" data sets that represent a person or a group in their entirety, even incomplete and ambiguous datasets can be used for valuable text analysis if the data issues are made transparent. Strategies to enhance data interpretation include **building sub-corpora** and considering **alternative scenarios** in a **comparative analysis**.



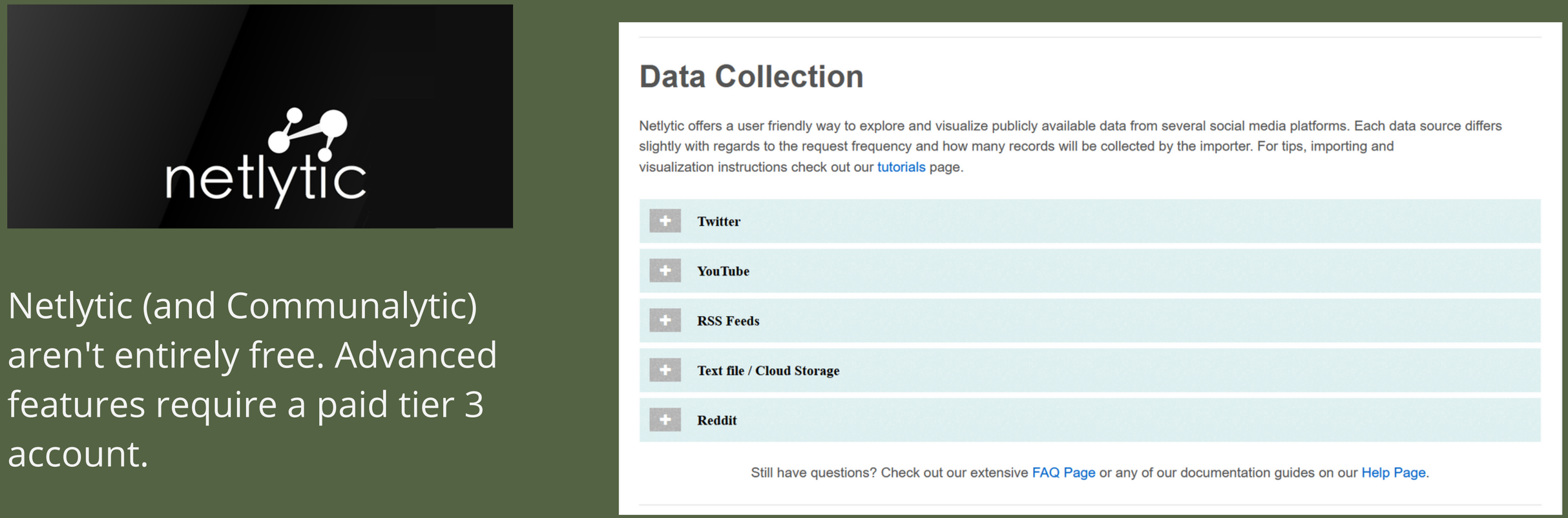
Tools for social media mining:

- Netlytic (user registration necessary)
- Web Data Research Assistant (Chrome browser extension)
- Communalytic.com (API restrictions apply)
- Facepager (Google)
- NodeXL (direct feed for Twitter but no longer free of charge)

RECOMMENDED TOOL I: NETLYTIC

Netlytic is a browser-based service that allows you to collect social media / web data without programming skills: https://netlytic.org/index.php

Netlytic (and Communalytic) aren't entirely free. Advanced features require a paid tier 3 account.



RECOMMENDED TOOL II:
WEB DATA RESEARCH ASSISTANT

The Web Data Research Assistant (aka WebDataRA)

WebDataRA is a tool has been developed by Prof Leslie Carr of the Web and Internet Science research group to support researchers using Web sourced data. Its functionality focuses on Twitter, but it also works for Facebook, Google, Google Scholar, SSRN, Github, Quora, ACM DL, Core and (since June 2020) Parler. The ultimate aim is to make advanced network analysis and textual analysis methods accessible to social science and non-programming researchers, especially those who work in an interdisciplinary context.

Installation information and documentation:
https://southampton.ac.uk/~lac/WebDataResearchAssistant/

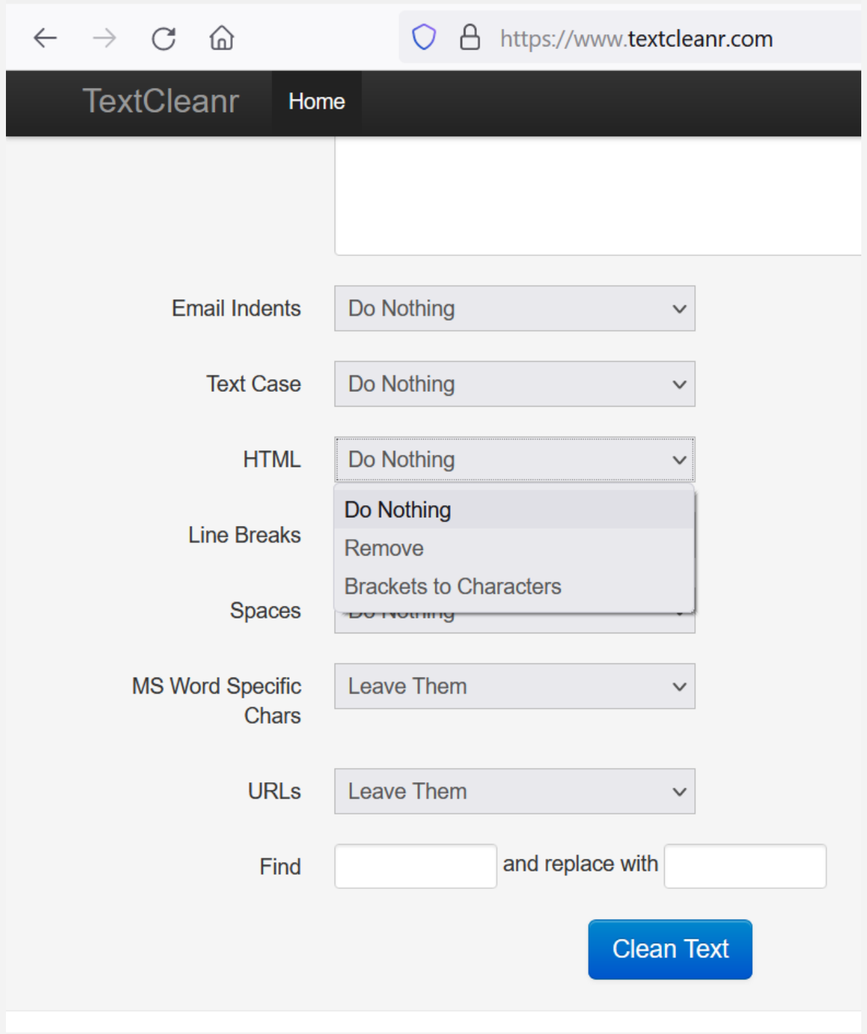
Text pre-processing / data cleaning

Common aspects of data cleaning for text analysis:

- focus on texts in one language (mostly part of corpus building)
- use stopwords
- get rid of symbols, mail addresses, time-stamps etc.
- get rid of punctuation
- ignore uppercase letters
- tokenization (necessary when analysing data with scripts)
- lemmatization (ignoring word inflections, widely applied in the humanities and social sciences)
- stemming (reducing words to root, mostly in linguistics)

TEXT CLEANR

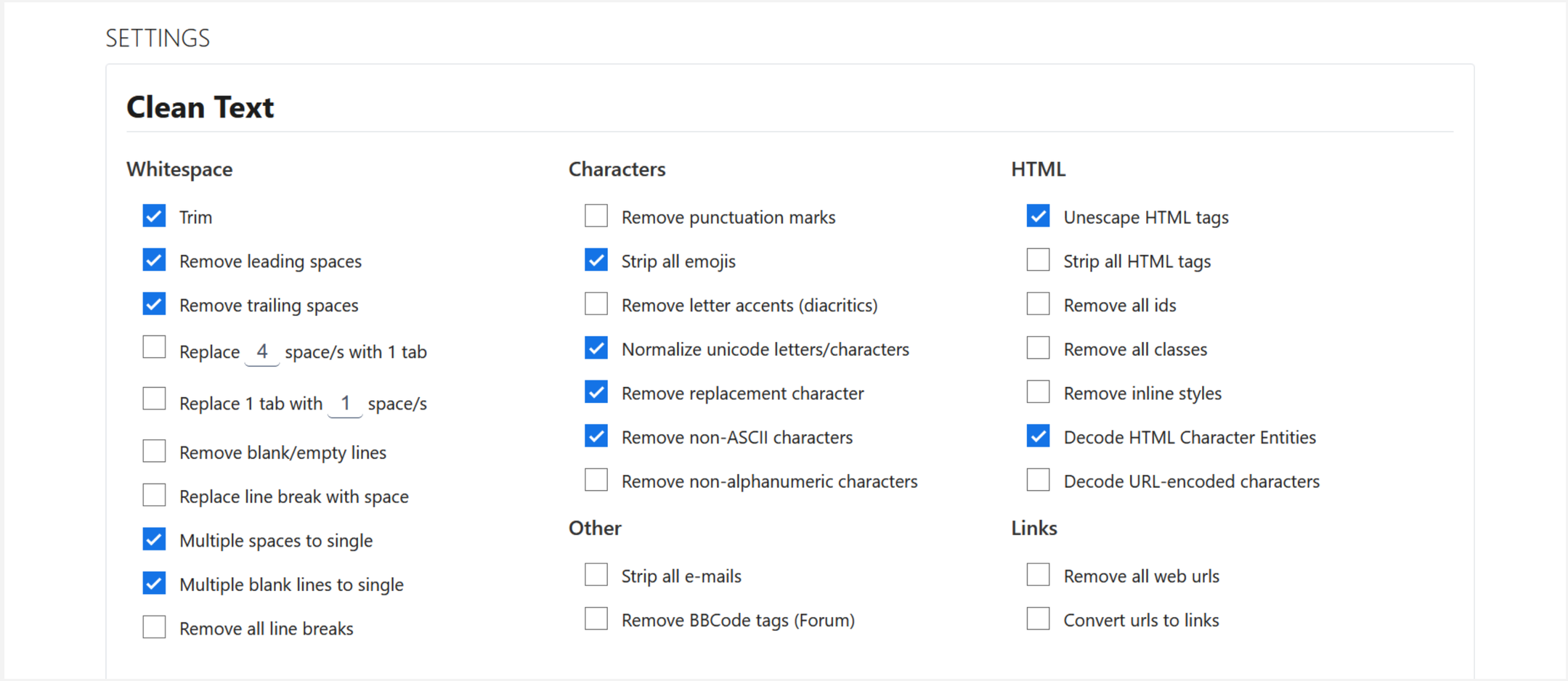
(https://www.textcleanr.com/)



- easy to use
- cleaning options in drop down menues
- limited cleaning possibilities (mainly for web-data)

TEXT CLEANER *net*

(https://textcleaner.net/)



Take a break...

(Visual) output of NLP operations

A lot of user-friendly distant reading tools place an emphasis on visualisations, but…

… creating graphs is **not** the primary objective of computational text analysis.

Graphs can support your research process and the communication of your results to different audiences.

IMPORTANT:

In "Machines of Knowledge", we ask you to analyse text mainly based on **visual representations**, but you ought to keep in mind that all NLP visualisations depend on **specific algorithms** and are not "accidental".

Algorithms in NLP:

NLP algorithms are based on machine learning, especially statistical machine learning. Some commonly used NLP algorithms include:

- Edit distance (also known as Levenshtein distance)
- Cosine similarity
- TF-IDF (term frequency and inverse document frequency)
- Naive Bayes algorithm (Naive Bayesian Analysis/NBA)

Types of graphs used in data/text analysis:

- column charts / bar graphs (comparisons between categories)
- histograms (continuous data measured on an interval scale)
- stacked bar / column charts
- area charts (comparing trends over time)
- dual axis chart (relationship between two variables)
- line graph (changes over time)
- mekko chart (development by segment)
- pie chart (percentages of a whole)
- waterfall chart (running total as values are added or subtracted)
- bubble chart (multi-variable)
- scatter plot (multi-variable)
- bullet graph (performance comparison)
- funnel chart (values across multiple stages in a process)
- heat map (values indicated by colour scales)
- tree map (hierarchies)
- chord diagrams (weigthed relationships)

Hands-on exercise I: use a word cloud for high-level text analysis

The following slide contains a word cloud in which the biggest words are the most frequent in a text. Use a search engine to find out which novel it is. What is the novel about? Why is that (not) reflected in the word cloud?

Example of a word cloud
(displaying the most frequent words in a text or corpus)

Hands-on exercise II:
create a word cloud with
https://www.wordclouds.com/
and compare your results

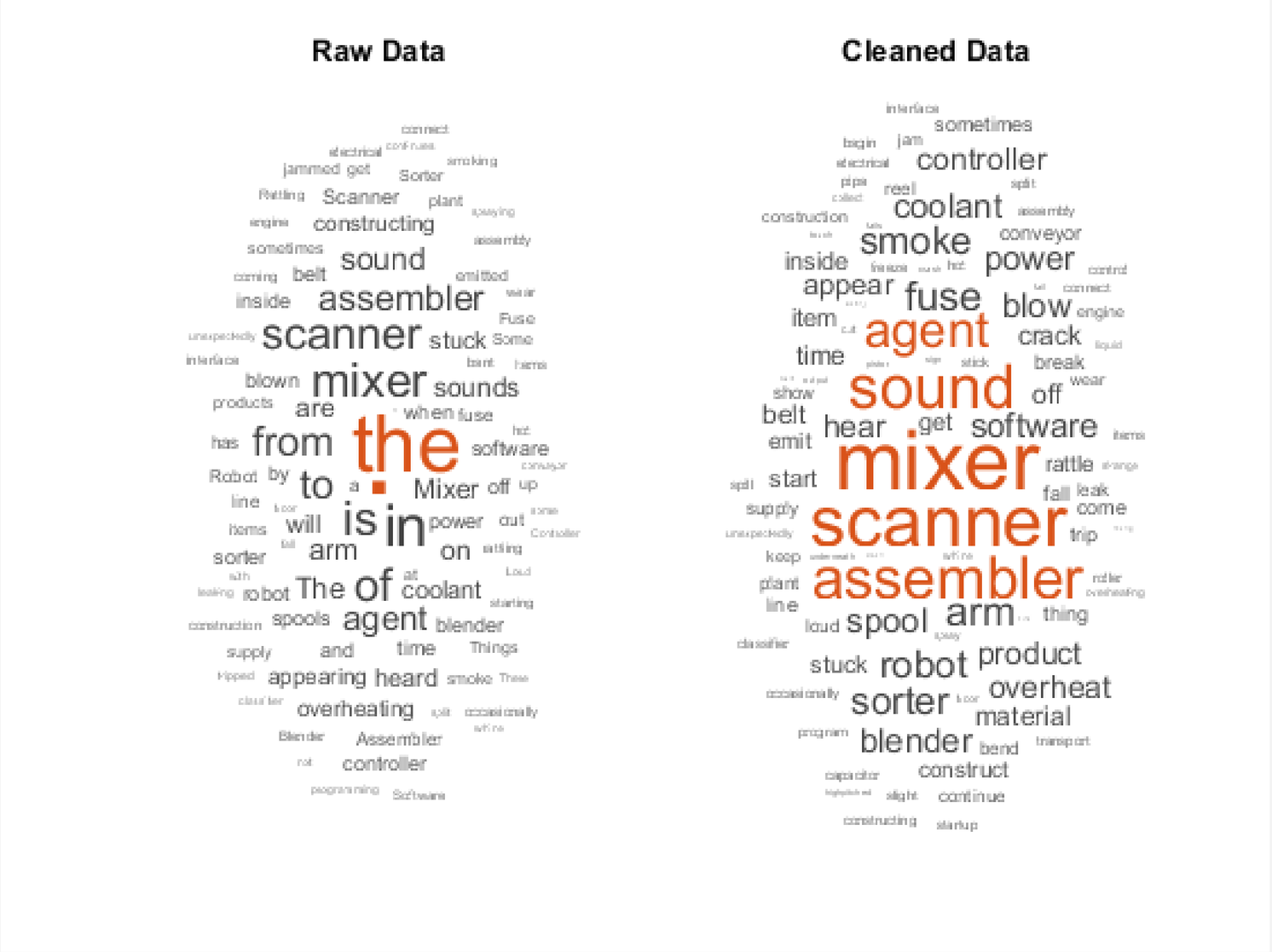
Text samples on GITHUB:

https://github.com/MonikaBarget/DistantReading/tree/main/YouTube_AmandaGorman

Differences between raw text and pre-processed text:

Why is the word cloud on the right "better"?

Example taken from https://se.mathworks.com/help/textanalytics/ug/prepare-text-data-for-analysis.html



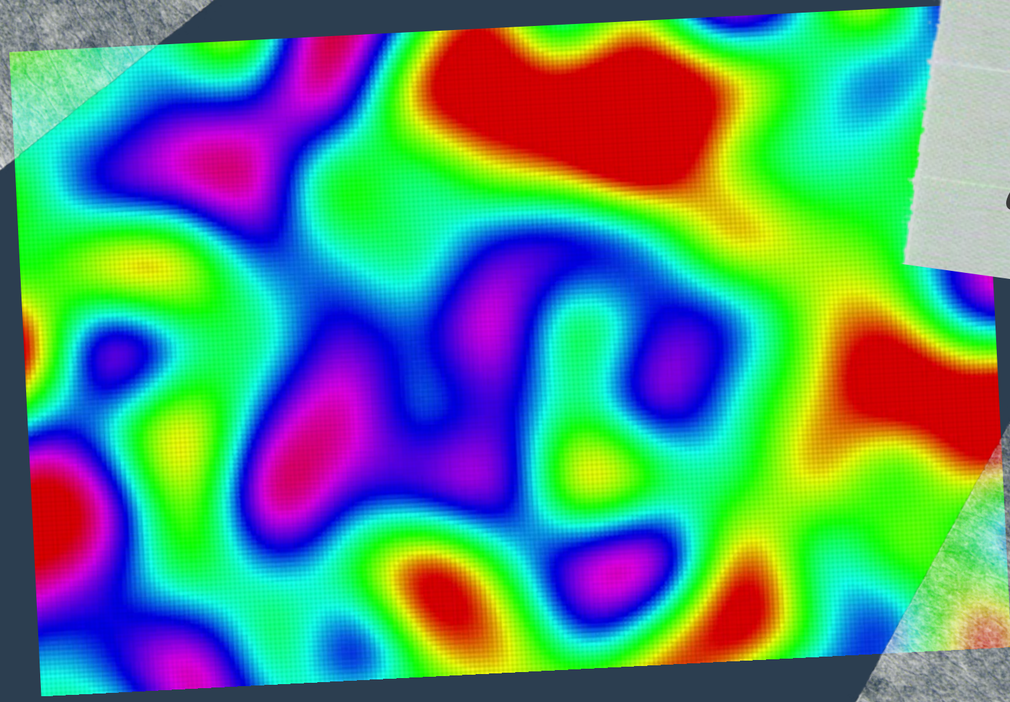
Creative, more meaningful uses of word clouds in research and teaching:

An alternative visualisation for the 2016 US "State of the Union" summary, created by Marti Hearst, 23 April 2019: https://medium.com/multiple-views-visualization-research-explained/improving-word-clouds-9d4a04b0722b

Hands-on exercise III: learn to interpret a "trends" line graph

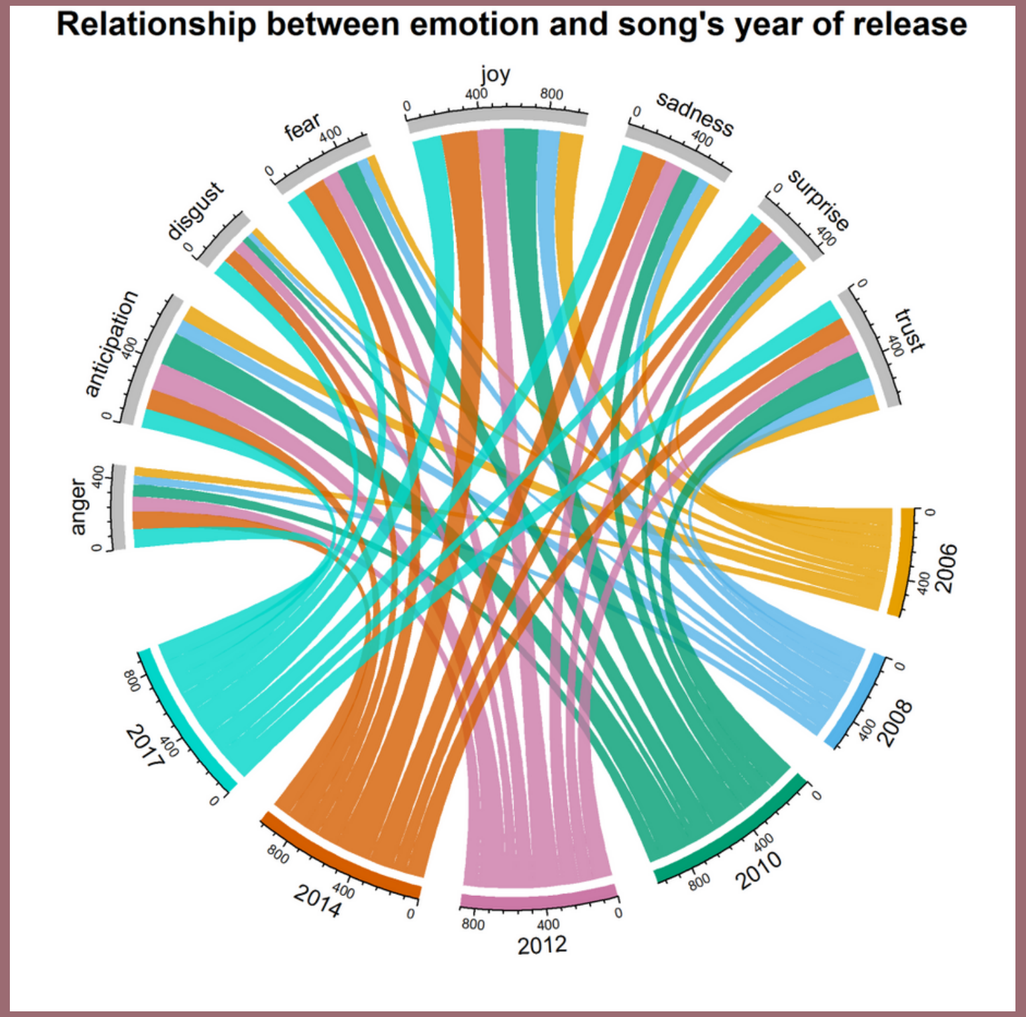
The following slide contains a line graph with several lines that cross each other. What do the individual lines stand for? What does the graph say about the analysed text?

Example of a line graph showing "trends" in a novel

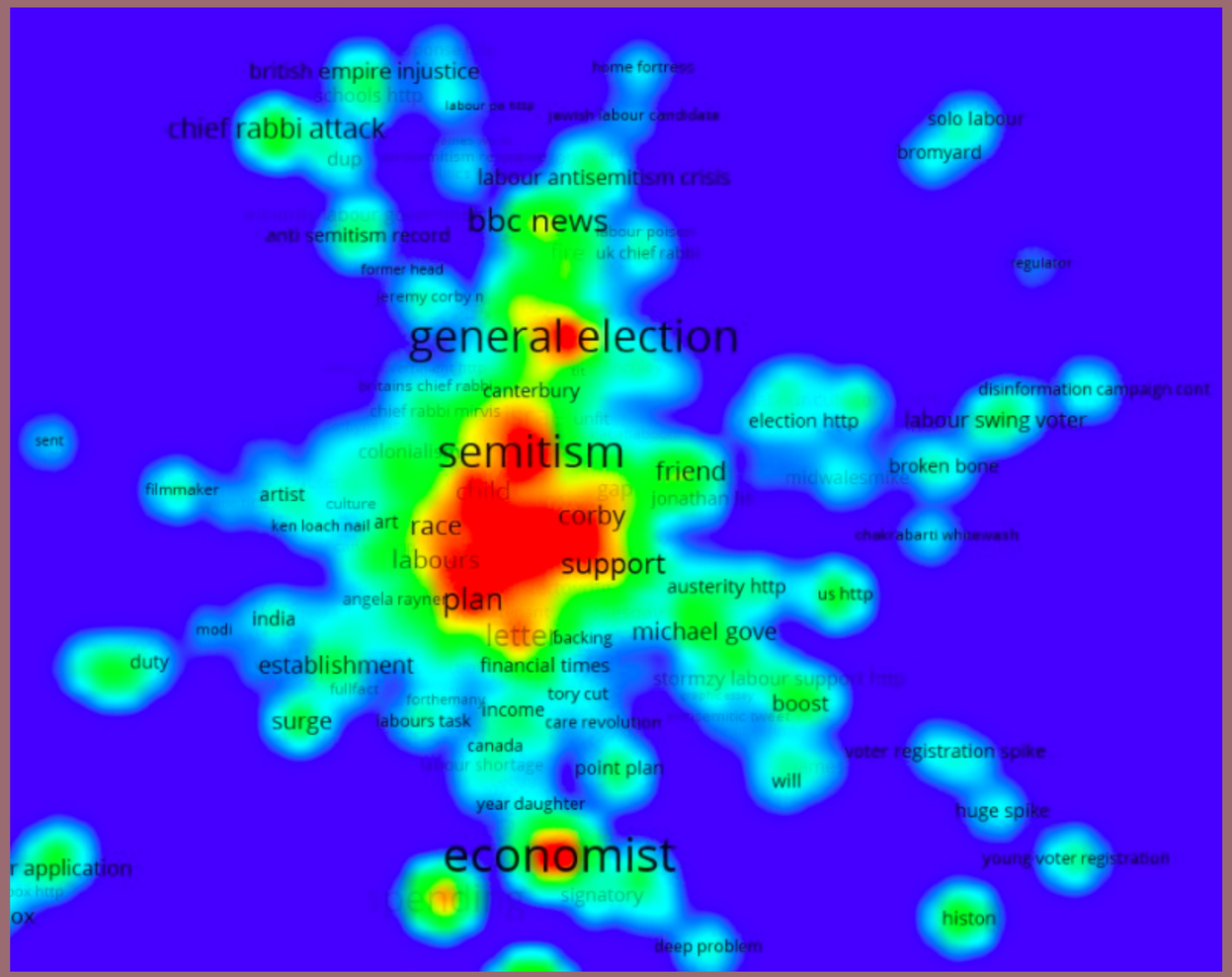


Hands-on exercise IV: identify the different graphs and discuss their usage in text analysis

Look at the two graphs on the following slide, go to the indicated websites, identify both use cases, and discuss the analytical value of the visualisations.



https://www.promptcloud.com/blog/data-visualization-text-mining-taylor-swift-song-lyrics/



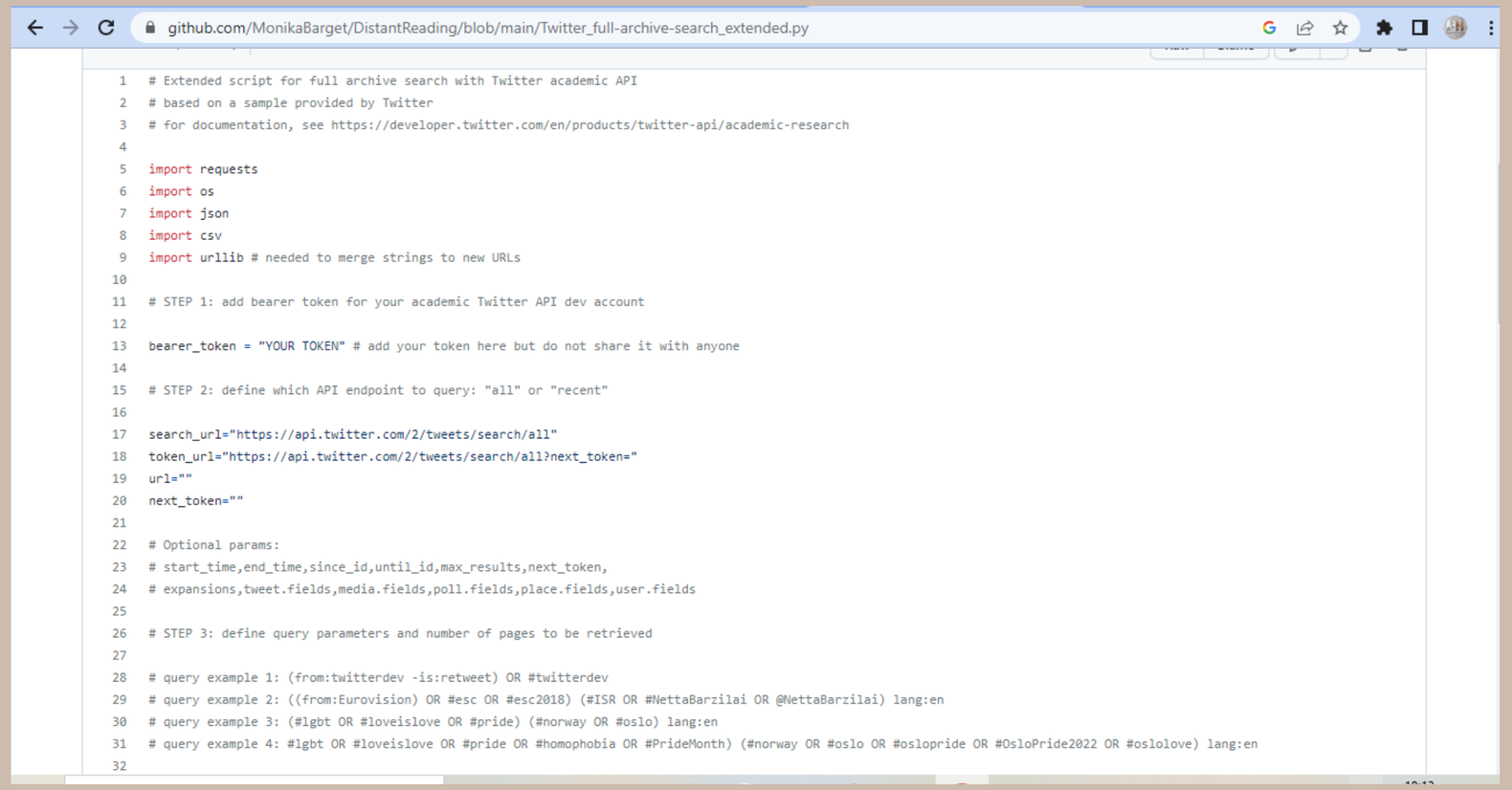
https://www.textminingsolutions.co.uk/news/hot-topics-on-a-heat-map/

How to perform text analysis...

"Python" and "R" packages:

Different programming languages such as Java and R are used for natural language processing. R is especially common in the social sciences and for linguistic analysis. Historians and other scholars with a content focus often favour Python because of its numerous packages, simple syntax, and its easy integration with other programming languages.

Sample script: extracting tweets with Python



https://github.com/MonikaBarget/DistantReading/blob/main/Twitter_full-archive-search_extended.py

Sample project: Irish Women's WWI Networks

FeministDH

Visualisations of women's correspondence in "Letters 1916-1923"

Women's agency and networks in Ireland (1915-1923)

The visualisations presented here were created by Monika Barget and Susan Schreibman, using metadata and transcriptions from the "Letters 1916-1923" project. Prof. Susan Schreibman (formerly Trinity College Dublin and NUI Maynooth, now Maastricht University) started the Letters 1916-1923 project in 2013. Since then, circa 6000 letters, telegrams, postcards and notes relating to Irish history between December 1915 and the end of the Civil War in 1923 have been collected from 36 private and 29 public archives.

Letters 1916–1923
Ordinary Lives - Extraordinary Times

The initial quantitative analysis of our metadata categories highlighted the gender ratio per manually assigned tag / topic in all items uploaded to the Letters 1916-1923 backend by December 2019. In this way, we could uncover themes that were [predominantly addressed by women] (https://github.com/MonikaBarget/FeministDH/blob/master/Letters_topics-covered-by-women.png).

Text analysis based on correspondence collected for the Letters 1916-1923 project, analysed by Susan Schreibman and Monika Barget using Python, https://monikabarget.github.io/FeministDH/

GUI tools for text analysis

BEGINNERS' TOOLS FOR DISTANT READING:

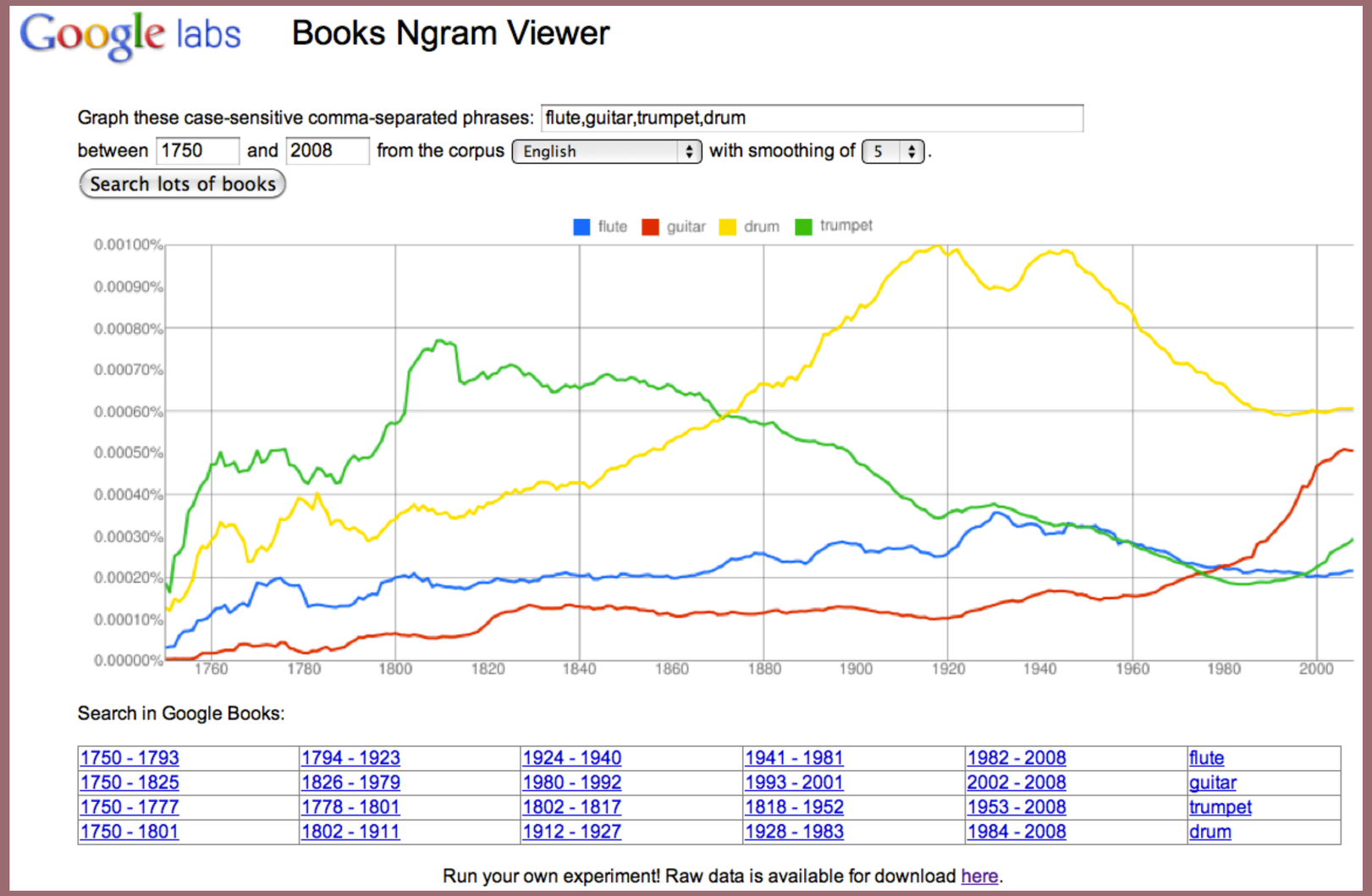
VOYANT

see through your text

Browser-based
tools
(e.g. word cloud generators,
Google n-gram viewer and Voyant
Tools)

Stand-alone software
with GUI
(e.g. DARIAH
Topics Explorer)

Tool example 1: Google n-gram viewer



Finding musical instruments in books published between 1750 and 2008:

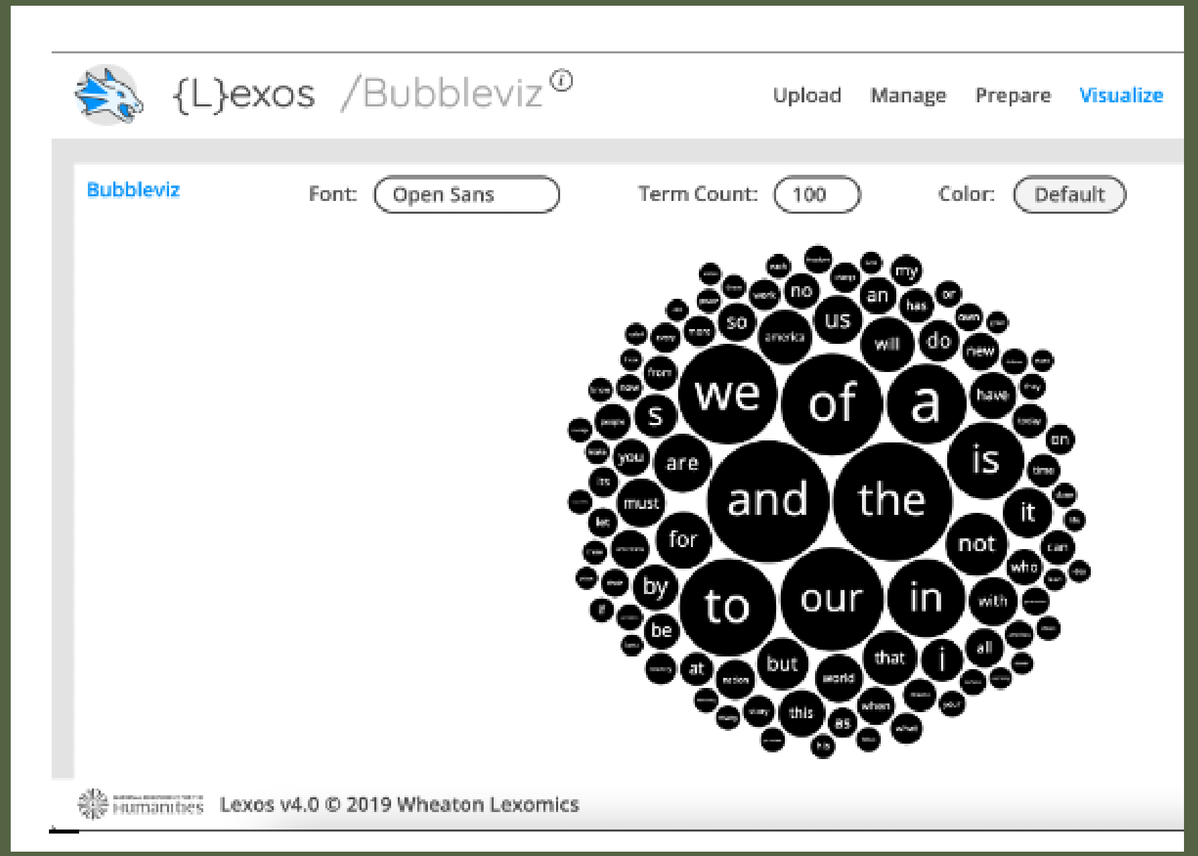
https://ai.googleblog.com/2010/12/find-out-whats-in-word-or-five-with.html

(Dis-)advantages of Google n-gram viewer

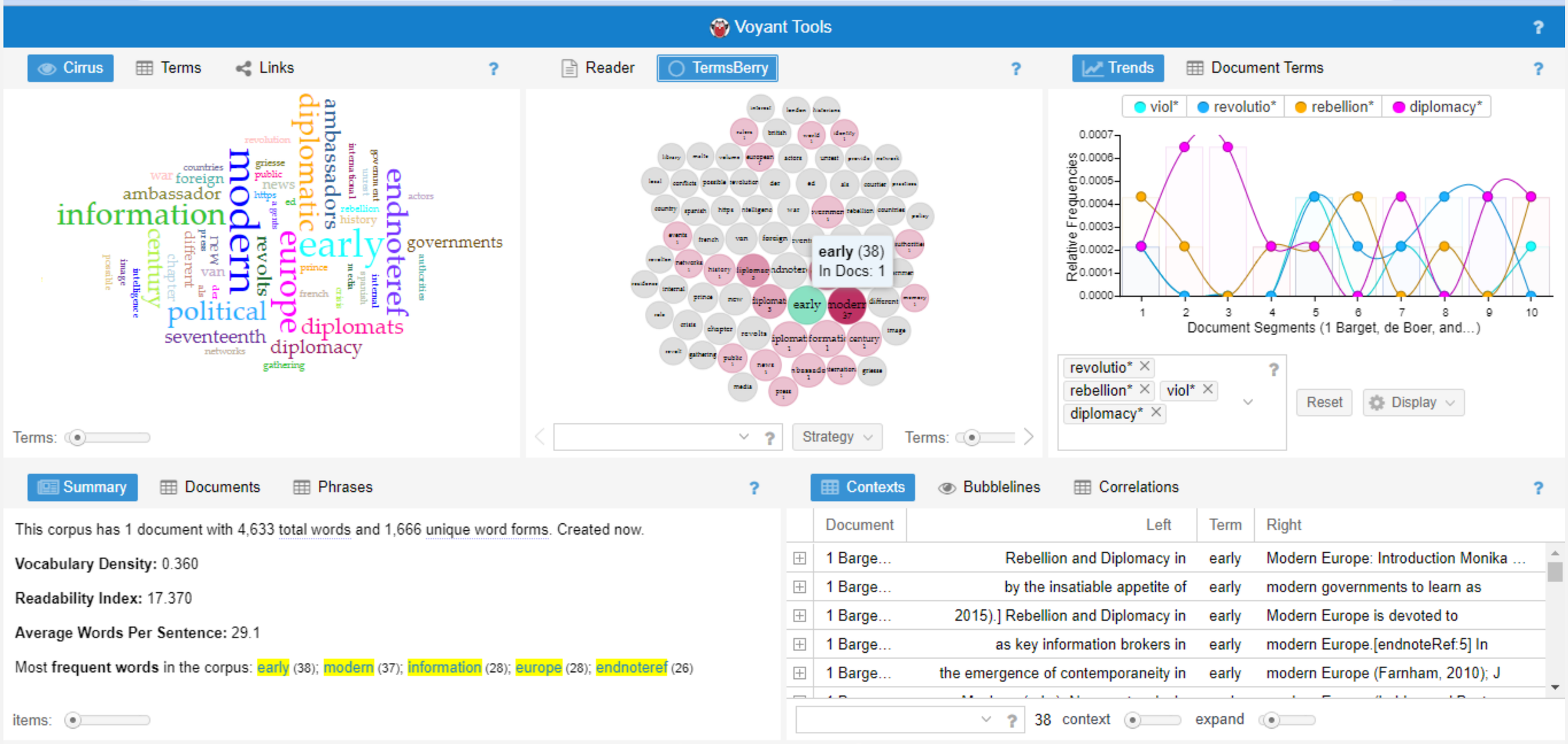
- easy way to graph and compare phrases from publications over time
- insights into long-term development of word usage
- large corpus of circa 500 billion words in total
- exciting accidental discoveries ("serendipity")
- no own corpus building
- corpus provided by GOOGLE intransparent and in progress
- limited algorithm control

Tool example 2: lexos.wheatoncollege.edu

- visualisations and statistics on word distribution
- option to scrape text from websites via URL
- clean interface and "info" buttons with technical / mathematical details
- several text-pre-processing options
- export of graphs as PNG or SVG
- browser-based version not suitable for large text corpora
- **BUT**: frequent connectivity issues



Tool example 3: Voyant Tools (NLP tool collection)



Tool example 3: Voyant Tools

Functionalities:

- ingest your own corpus (via URL or file import)
- handle multiple languages
- explore text without programming skills
- use several tools for analysis and visualisation
- quickly change settings
- good documentation and extensive tutorials

Limitations:

- not always clear what tools are best suited to the imported corpus: some depend on multiple documents, others don't work well with hundreds of documents
- interface dense and at first overwhelming
- occasionally prolonged text-loading time

Customizable settings in Voyant:

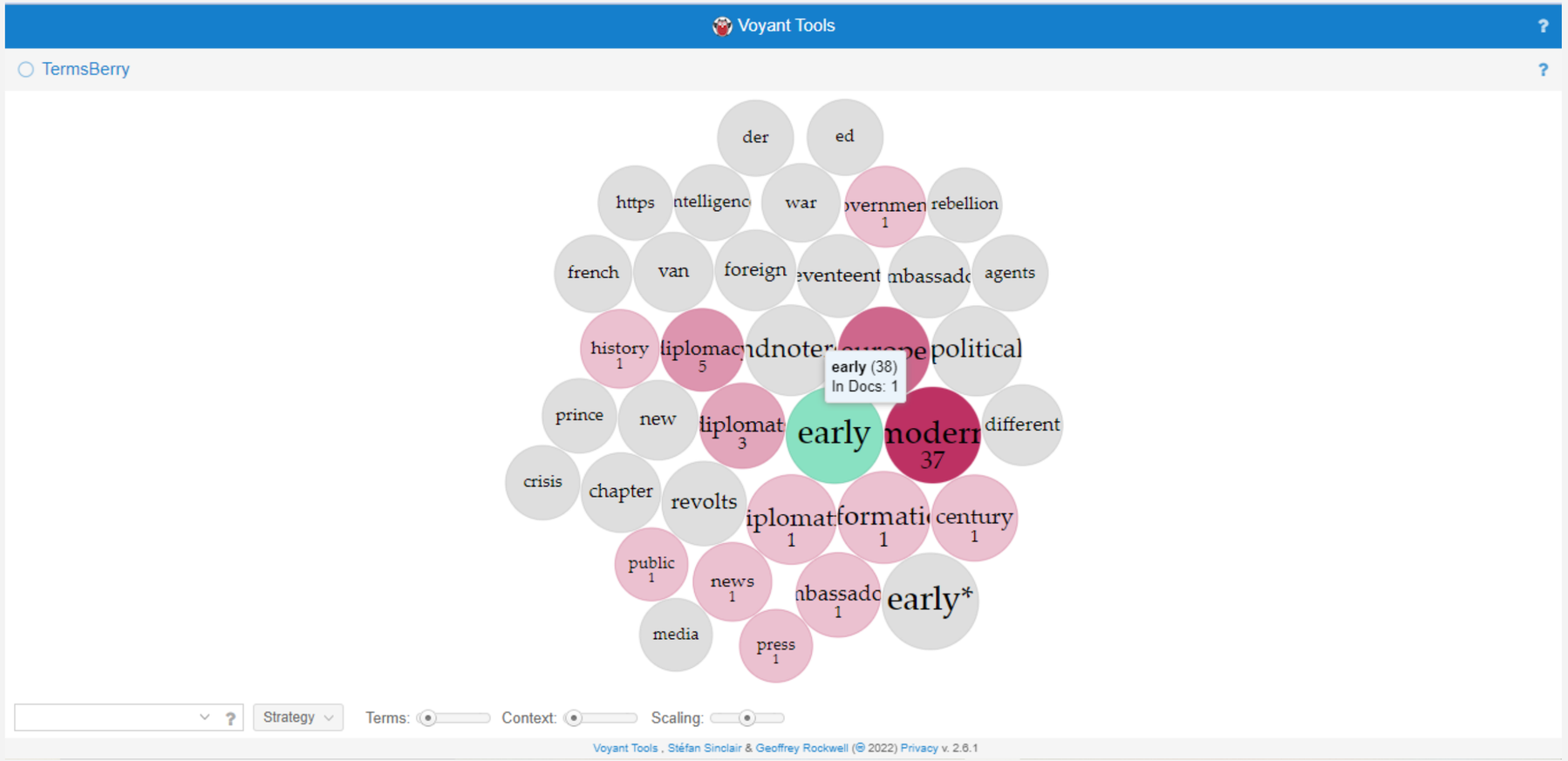
- turn stopwords on / off
- chose number of words to be displayed
- change layout of tool display (e.g. font size)
- adjust CONTEXT and CORRELATIONS
- export visualizations

We will cover Voyant in more detail in the 2nd skills session.
Watch out for the Voyant Tools tutorials recommended on CANVAS!

Voyant Tools "terms berry"

The TermsBerry tool shows high frequency terms in combination with information how those terms co-occur.

The **term frequency–inverse document frequency value** is a weighting factor that balances the overall frequency of words and their importance to individual documents.



Example: In a text on "diplomacy" in "early modern Europe", it is no surprise that "early" mostly co-occurs with "modern".

What's next?

What we will cover in the two up-coming skills sessions:

- detailed text analysis with Voyant Tools
- in preparation: "play" with the sample data (provided on GITHUB) at home
- understanding differences between visualisation options in Voyant
- working with Voyant table tools

Thanks for your participation!

m.barget@maastrichtuniversity.nl