Machines of Knowledge - Maastricht University - Period 2

Corpus Building & Text Cleaning

Selecting texts for analysis: challenges and tools

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?

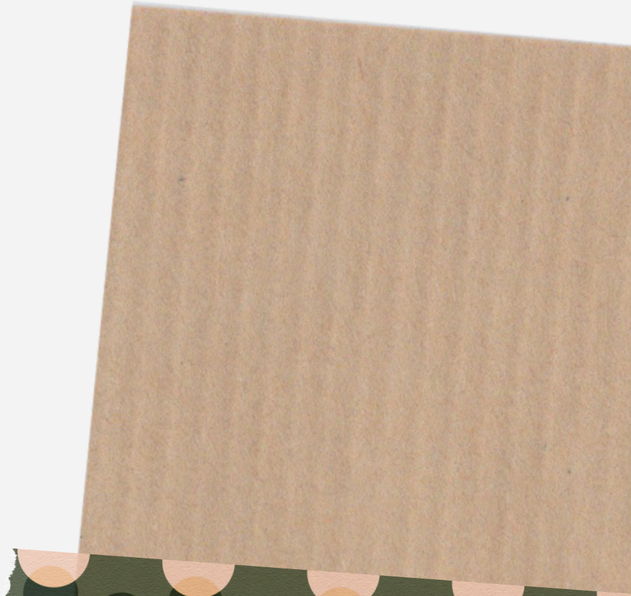
- "importance" of individual texts negotiable
- access to texts limited
- texts incomplete or lost
- human and computational ability to process multiple languages limited

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)

SELECTION CRITERIA

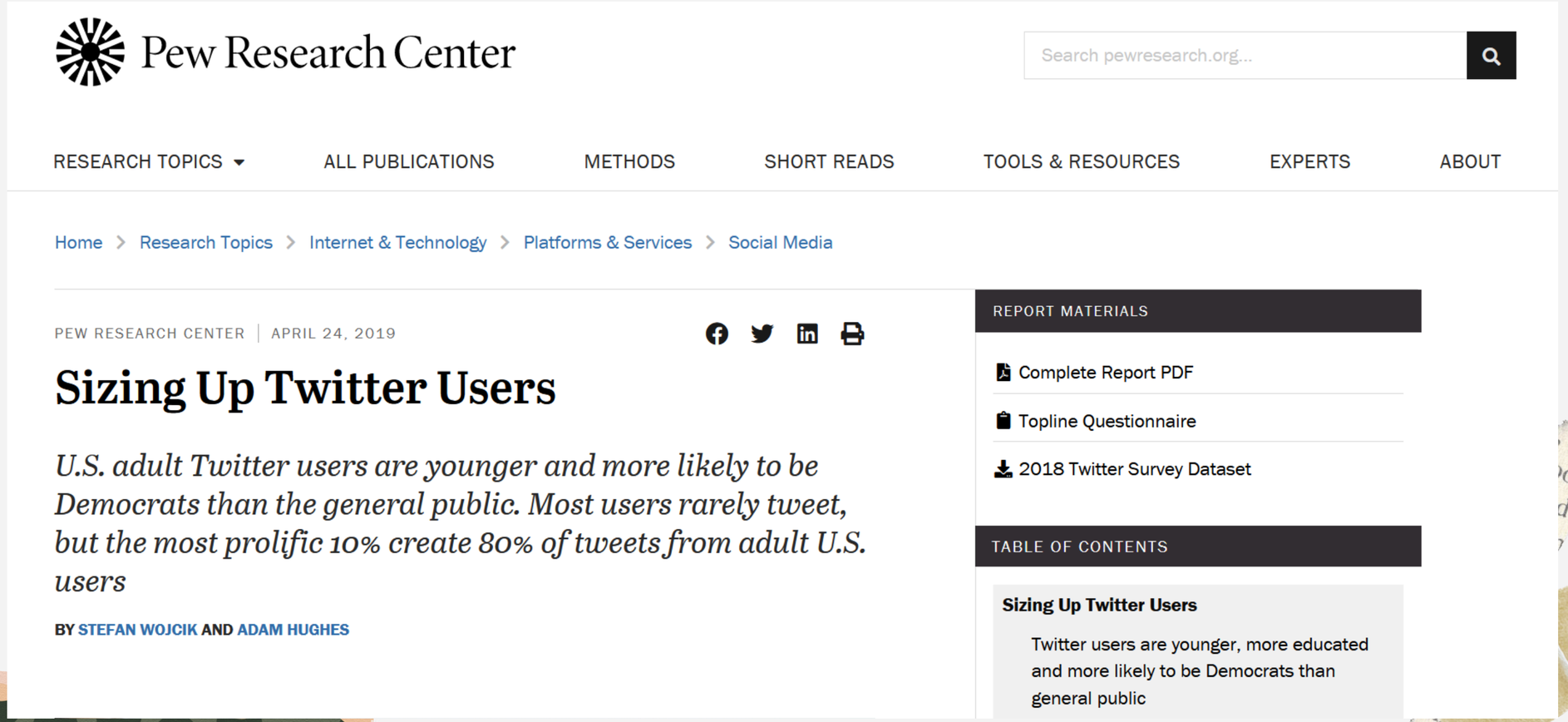
- topics / content?
- max. amount of data available?
- timeframe / points in time?
- people / demographics involved?
- comparison between different genres, authors, or audiences?
- geographic aspects?
- languages?
- gender, race, other categories?



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**.

Browser-based tools for social media mining:

- Netlytic (more advanced access needs subscription)
- Web Data Research Assistant (academic)
- Communalytic.com (API restrictions apply)
- Facepager (Google)

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.

WEB DATA RESEARCH ASSISTANT: https://southampton.ac.uk/~lac/WebDataResearchAssistant/

NETLYTIC: https://netlytic.org/index.php

"Python" and "R" packages to access (social media) data with scripts:

- API access to harvest Apple reviews
- API access to collect YouTube comments
- note that access rules can change (see Twitter / X after Musk) so check current regulations carefully
- scraping content via HTML is potentially illegal and not recommended in research

R Programming

Building a corpus with Python

DistantReading Data Collection Data Cleaning Data Analysis Samples Datasets Intro Data Scraping NLP Skills T

Distant Reading
a repository for teaching

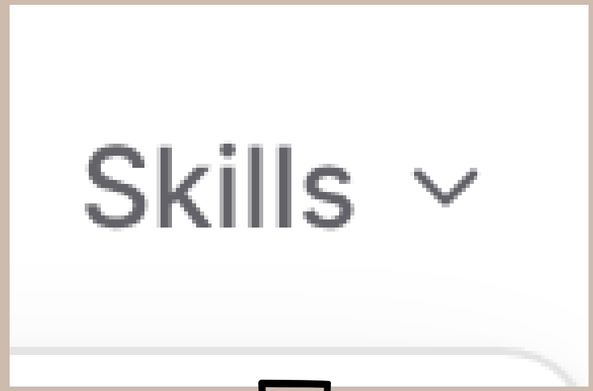
Welcome | Github repository | Jupyterlite

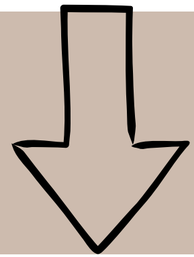
📊 Data collection
Learn how to scrape data for text analysis using Python scripts.

🧹 Data cleaning
Find out how to clean and filter data with EXCEL, Open Refine, scripts and browser-based tools.

📈 Data analysis
Learn how to use Voyant Tools for computational text analysis.

📺 YouTube channel
Visit YouTube to find examples of code execution and tools tutorials.





Getting Started with JupyterHub on DSRI (DCU4008)

In this course, we will use **Python**, a popular programming language for data analysis, to scrape and clean data from the web. Python is widely used in research and industry due to its simplicity and powerful libraries (collections of reusable code). To run Python code, we use **Jupyter Notebooks**, an interactive and user-friendly environment that combines code, visualizations, and explanatory text in a single document.
Students who have completed the BA Digital Society degree at FASoS may already be familiar with Jupyter Notebooks, but even if you have never seen it before, you will quickly learn how to use them.

Take a break...

Text pre-processing / data cleaning

DATA BIASES & SAMPLING METHODS

SAMPLING OPTIONS AND THEIR POTENTIAL PROBLEMS:

- time samples (day by day)
- pre-selection by keywords and hashtags
- by user names (e.g. assuming gender)

Recommended reading:

Kim H, Jang SM, Kim S-H, Wan A. Evaluating Sampling Methods for Content Analysis of Twitter Data. Social Media + Society. April 2018. https://journals.sagepub.com/doi/10.1177/2056305118772836

DATA WRANGLING / PRE-PROCESSING

"A good data wrangler knows how to integrate information from multiple data sources, solve common transformation problems, and resolve data cleansing and quality issues."

Divya Naidu, POS Database Analyst at MINA Group | Research Assistant at Rutgers University, June 2020: https://towardsdatascience.com/twitter-analytics-weratedogs-a441be7d4a85

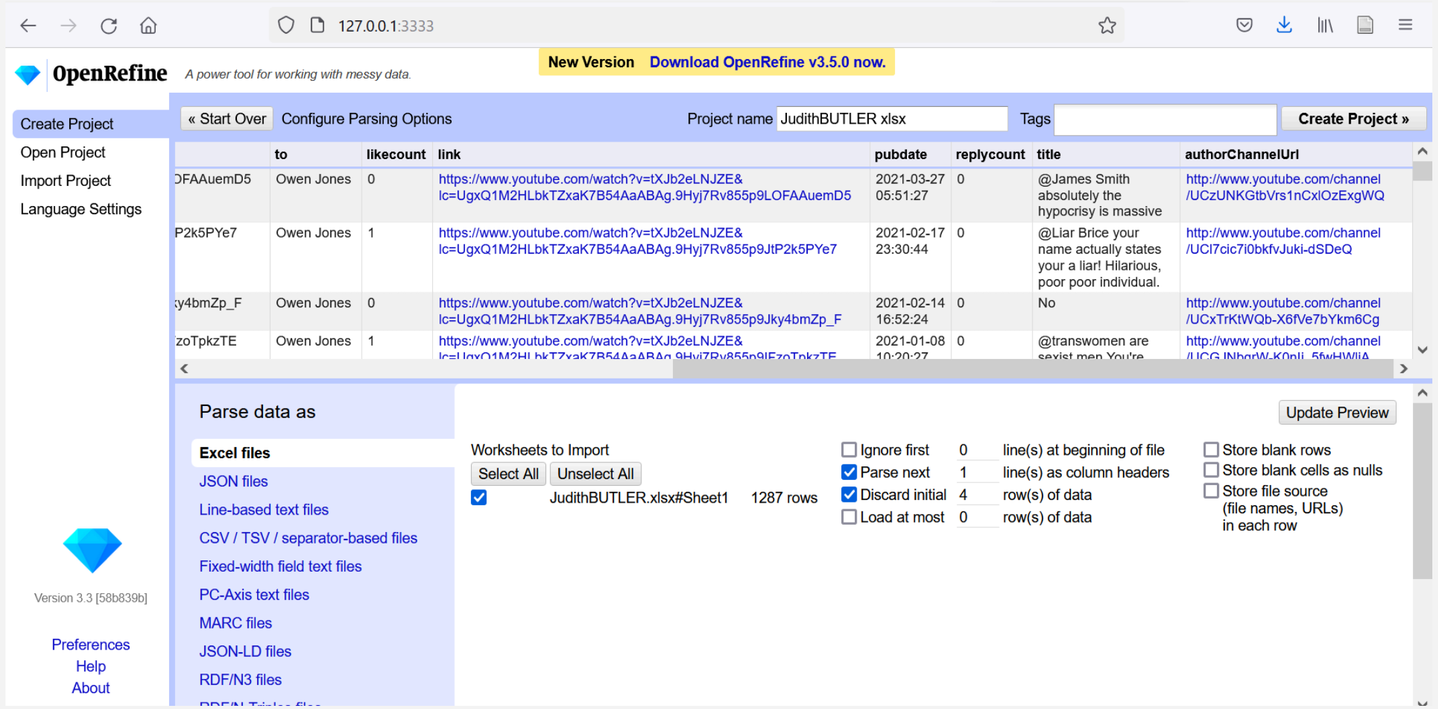
Common aspects of data cleaning for text analysis:

- 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)

Tools for data cleaning:

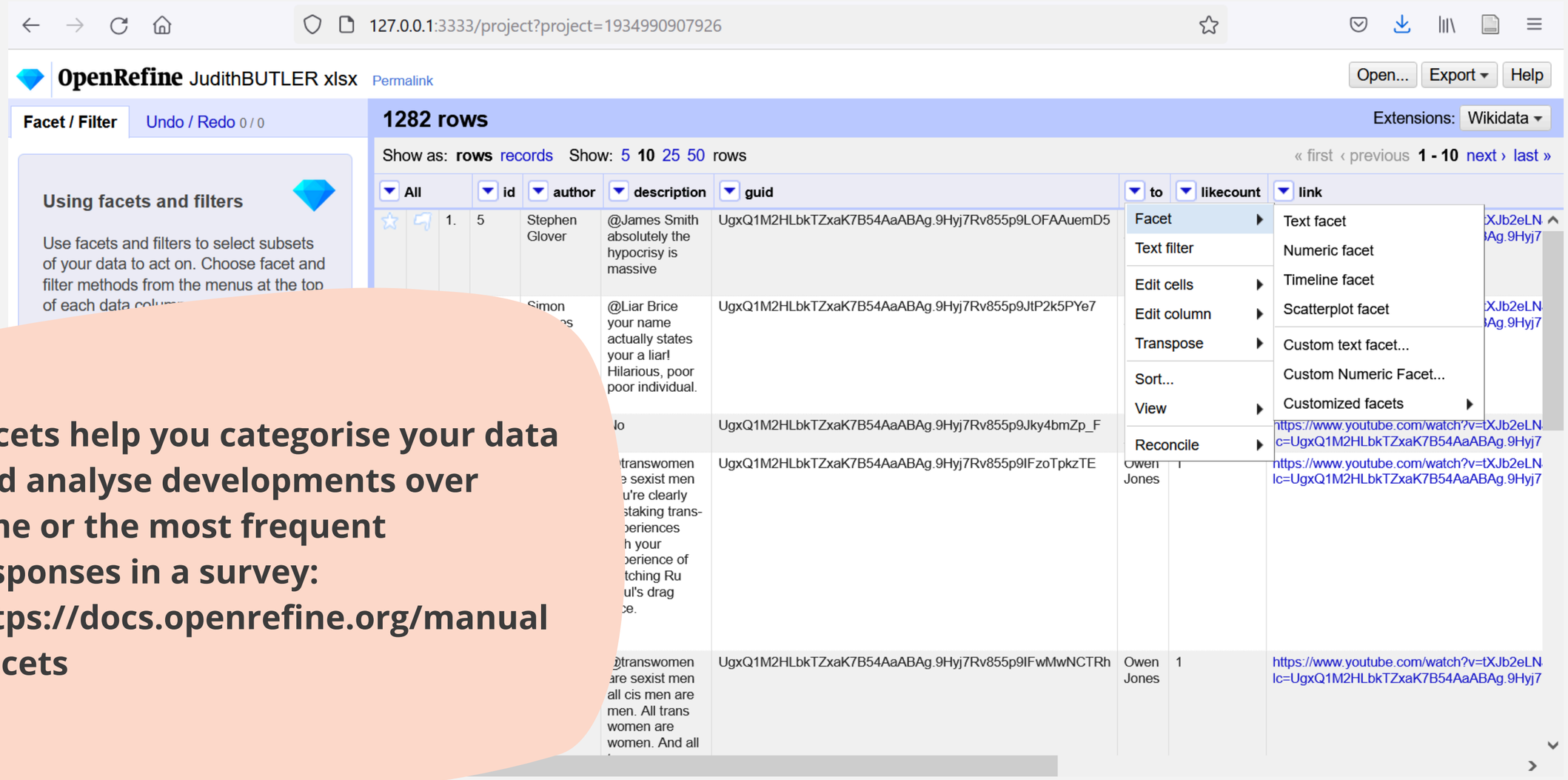
- TextCleanr (web-based)
- Text Cleaner (web-based)
- Open Refine (installation)
- Trifacta Wrangler (cloud)
- TIBCO Clarity
- Data Ladder (installation)
- Data Cleaner (installation)
- GOOGLE sheets (browser)
- etc.

Cleaning data tables in OPEN REFINE *(software installation needed):*



- define data headers
- ignore columns / rows with unwanted data or no data
- "create project" to carry out more detailed cleaning
- use facets and filters to select subsets of your data
- facet and filter methods are available at the top of each column

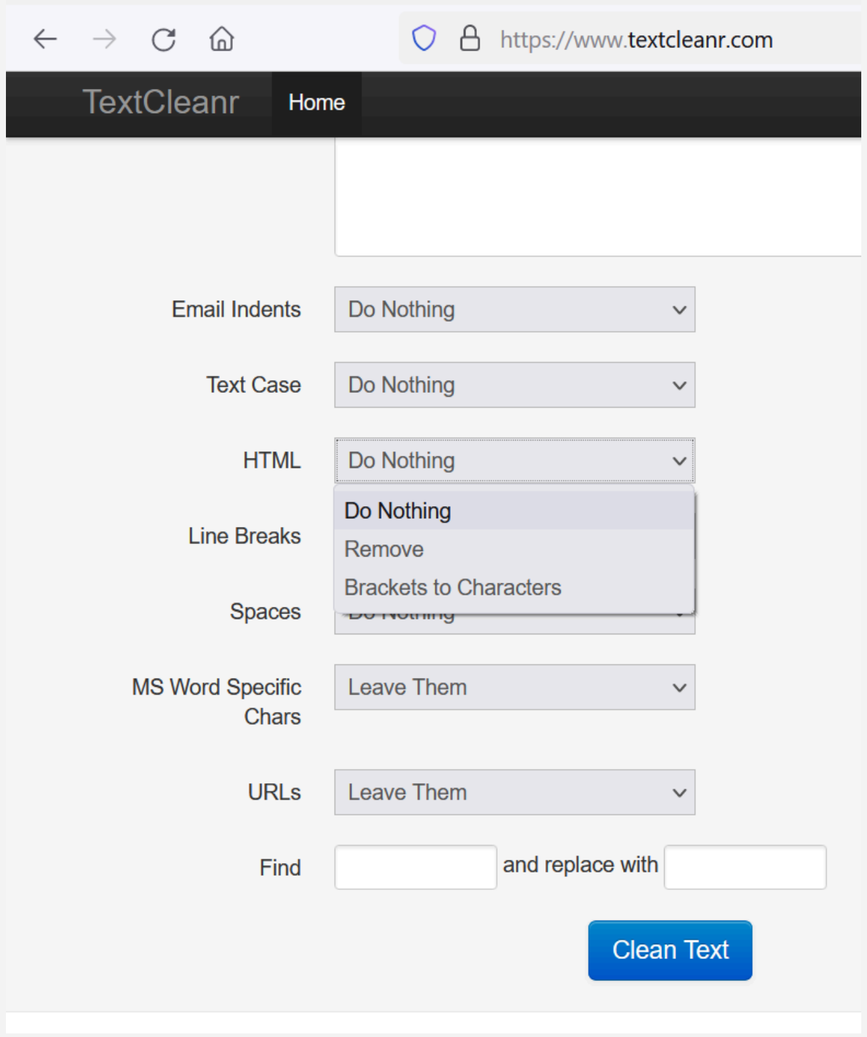
Defining facets for OPEN REFINE *data columns:*



Facets help you categorise your data and analyse developments over time or the most frequent responses in a survey: https://docs.openrefine.org/manual/facets

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/)

SETTINGS

Clean Text

Whitespace

- [x] Trim
- [x] Remove leading spaces
- [x] Remove trailing spaces
- [ ] Replace 4 space/s with 1 tab
- [ ] Replace 1 tab with 1 space/s
- [ ] Remove blank/empty lines
- [ ] Replace line break with space
- [x] Multiple spaces to single
- [x] Multiple blank lines to single
- [ ] Remove all line breaks

Characters

- [ ] Remove punctuation marks
- [x] Strip all emojis
- [ ] Remove letter accents (diacritics)
- [x] Normalize unicode letters/characters
- [x] Remove replacement character
- [x] Remove non-ASCII characters
- [ ] Remove non-alphanumeric characters

Other

- [ ] Strip all e-mails
- [ ] Remove BBCode tags (Forum)

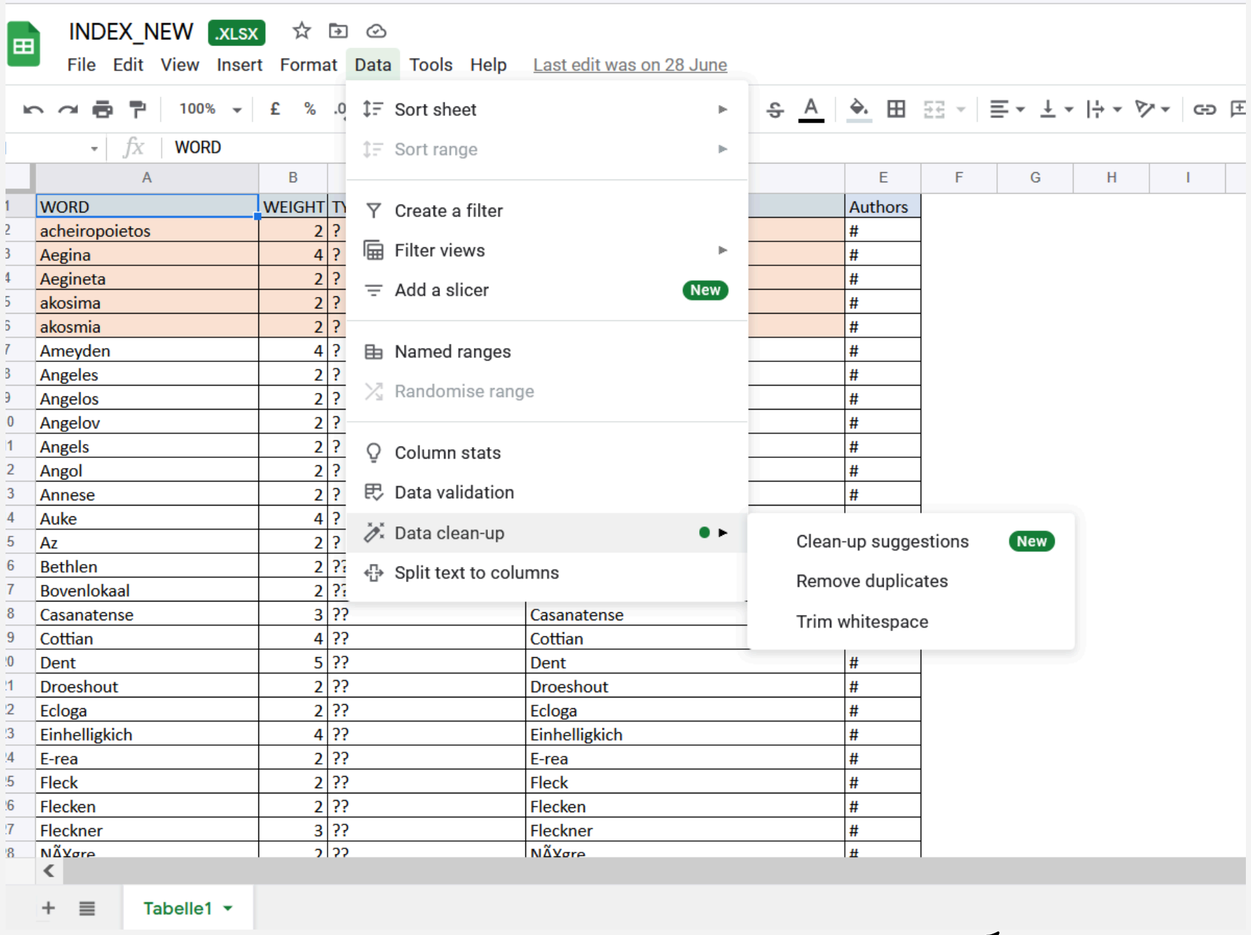
HTML

- [x] Unescape HTML tags
- [ ] Strip all HTML tags
- [ ] Remove all ids
- [ ] Remove all classes
- [ ] Remove inline styles
- [x] Decode HTML Character Entities
- [ ] Decode URL-encoded characters

Links

- [ ] Remove all web urls
- [ ] Convert urls to links

Cleaning data tables in Excel *or* GOOGLE SHEETS:



GOOGE SHEETS -> *Data -> data clean-up -> Clean-up suggestions*

- split data into columns (by delimiters or tab)
- conditionally delete rows (e.g. by keyword)
- swap rows and columns
- remove duplicate data
- remove extra spaces
- choose “capitalization” from the "format" menu
- use keyboard shortcuts (grouping or hiding data)

DATA FORMATS AND FILE NAMES:

- file formats depend on text analysis tool you wish to use
- file formats recognised by Voyant: HTML, MS Word, MS Excel, ODT, Pages (Apple), PDF, plain text, RTF, XML, archives (.zip, .tar, .tgz, etc.)
- however: distinction between **data** and **metadata** can be difficult
- ingesting **necessary data only** is recommended
- file names should follow technical recommendations (ideally no spaces and special characters) but also be human-readable
- use short und clear file names for a corpus of many texts, e.g. **[author]+[date].txt**
- save all files you wish to analyse in one directory

Thanks for your participation!

m.barget@maastrichtuniversity.nl