Text cleaning with SAGE Texti

Release V1

SAGE Publishing

Sep 22, 2021

FIND MORE:

CHAPTER

ONE

HOW TEXTI WORKS?

Texti is designed to help you find the best way to clean and preprocess your text documents. It currently supports pdfs only, and works on single files. You can preview how your chosen transformations extract and process the text from your pdf. Once you are happy with the output, you can download the python code behind the sequence in a jupyter-formatted notebook.

First, create a new project, give it a name, and a bit of an explanation of what you intend to do or which corpora would this most likely apply to.

Second, you will upload a pdf file, pick any from your working corpus.

Once you uploaded the file, you can go crazy on the transformations. Feel free to select and try all of them.

Finally, after you picked and played around with the transformations and are happy with what you see, you can download the python code. You might need to edit it slightly before you can run it on your entire corpus.

Note: Do you need more help, or have ideas to improve Texti? Contact daniela (dot) duca (at) sagepub.co.uk

CHAPTER

TWO

TRANSFORMATIONS

Texti supports several different text preprocessing transformations that you can mix and match for your specific needs. You can also create workflows or sequences of transformations. The table below summarizes and provides examples of current transformations in Texti.

If you're interested in adding to the list, consider contributing! More details in the contribute page.

CHAPTER

THREE

CORPORA

This page contains several corpora relevant to political science research, categorized by country and key source, a link for where to find them and a note if they are not free. We are working with many on these to develop Texti.

3.1 Parties and elections

Item	Country	Description	Access	Link
Manifesto Project	51 inc. OECD	All political manifestos from the first democratic election onwards.	API; stata, spss, csv, xslx	Here
Speeches	UK	Speeches from party leaders from 1895 to today	HTML on site	Here
Regional manifesto	Spain	1980 to 2019, all regional parties	Download	Here
Regional manifesto	Wales and Scotland	1999 and 2016	Download	Here
Regional manifesto	Italy	Fragmented depending on region	Download	Here

3.2 Parliament Activity

Item	Country	Description	Access	Link
Parliamentary Questions Answered	UK	278428 questions; csv	API	Here
EP Plenary	European Union	1997 to 2019	HTTP resolvable URIs	Here
Parliament Debates	France	Debates of the 15th legislature	HTTP resolvable URIs; XML	Here
Lords Written Questions	UK	52004 questions	API; csv	Here
Commons Written Questions	UK	275929 questions	API; csv	Here
Questions to the Government	France	Since 2017	HTTP resolvable URIs	Here
Questions to the Government - without debates	France	Since 2017	HTTP resolvable URIs	Here
Written quesions to the Government	France	Since 2017	HTTP resolvable URIs	Here
Parliamentary Debates on Europe	France	2002 to 2012	HTTP resolvable URIs	Here
Parliamentary speeches	Austria, Czech Republic, Germany, Denmark, Netherlands, NZ, Spain, Sweden, UK, Ireland	21 to 32 years of data	API on DataVerse; full-text vectors in rds	Here
Parliament Rules	UK	1811 to 2019	Download	Here
Parliament Rules	Ireland	1922 to 2020	Download	Here
Debates and Replies to Questions	Ireland	All	API	Here
Senate “Dossiers Legislatifs”	France	Documents discussed since 1977	Download	Here
Amendments by the Senate	France	Amendments since 2001	Download	Here
Lords Bill Amendments	UK	11727 Amendments	API	Here
Questions to the Government (Senate)	France	Since 1978	Download	Here
Research Briefings	UK	9739 briefings	API, csv with 500 records limit	Here
Proceedings	European union	1996-2011	Download, xml	Here

3.3 Legislative Documents

Item	Country	Description	Access	Link
All legislation	European Union	Summaries of EU legislation (full corpus exists but wrong license)	HTML on site (can email Dimiter Toshkov for `Python` script)	Here
Trade agreements	European Union	All free trade agreements	List of linked PDFs	Here
Bills	UK	All bills since 2007	API	Here
All Legal Texts	France	Constitution, laws and decrees, court rulings, treaties (in French and translated)	Downloadable + beta API	Here
Legislation	Wales	All Bills, Acts, Marshalled lists	XML export	Here
The Record of Proceedings	Wales	All proceedings	XML export	Here
International Environment Agency	World	Most environmental treaties and agreements	List of .txt on the website	Here
Bills and Acts	Ireland	All	API	Here
All trade agreements	All	All	Download	Here

3.4 Identity and Culture

Item	Country	Description	Access	Link
National Anthems	World	194 countries	Download	Here

3.5 Presidential & Governmental Activity

Item	Country	Description	Access	Link
Political speeches	UK	8000+ political speeches on British Politics	HTML	Here
Official correspondence	UK	All official correspondence of PMs	API	Here
PM transcripts	Australia	Ministerial transcripts from 1940s to date	API; xml	Here
Speeches	EU	All ECB President / VP speeches	Download; csv	Here
Speeches	Germany	6,685 speeches by 71 officials, spanning a time from 1984 to 2017	Download, xml	Here
Speeches	EU	18,403 speeches from EU leaders from 2007 to 2015	API from DataVerse; csv raw speeches, and term-document matrices in R	Here
State of the Nation	South Africa	1990 to 2018	Download from Kaggle; txt per speech	Here

3.6 Participative democracy

Item	Country	Description	Access	Link
Public consultations	France	Recent public consultations	HTTP-resolvable URIs	Here
E-petitions	UK	All official e-petitions	API; JSON, xml, csv, HTML	Here

3.7 News and Media

Item	Country	Description	Access	Link
EUvsDisinfo	Europe	Debunked news articles by European External Action Services	API; HTML	Here
New York Times	All	Archive metadata, books, comments, reviews, most popular articles	API; JSON	e.g. Here
Public debates over European integration	Austria, Britain, France, Germany, Sweden, and Switzerland	1970s to 2012 from newspapers	csv, dta	Here
Public debates over globalization issues	Austria, Britain, France, Germany, the Netherlands, and Switzerland	2004-2006 from newspapers	csv, dta	Here
Archive of Political emais	Australie, Canada, France, Germany, Ireland, Italy, NZ, UK, USA	348,680 emails	HTML	Here
News articles	Not specified	9+ million articles and metadata for each	CSV split in 1GB zip files, download from GitHub	Here
Poliwoops	Many countries including USA, UK and most European countries	Deleted tweets by public officials and politicians	API; JSON	Here

3.7.1 Messy list of promising websites

Websites that might be goldmines but would require some time to explore.

- European Language Resource Coordincation
  - A lot of legal / official documents translated and sometimes already processed. E.g. IP case law, audits, a lot of legal texts from EU countries (not sure how useful they really are, but it is a *lot* of them, there might be some interesting ones)
  - https://elrc-share.eu
- Clarin
  - List of 24 parliamentary corpora, not all easy access
  - https://www.clarin.eu/resource-families/parliamentary-corpora
- EveryCRSReport.com
  - Reports from the Congressional Research Service — essentially the national legislature's think-tank.

  - https://www.everycrsreport.com/
- Supreme court transcripts
  - https://www.oyez.org/

3.7.2 Complementary text data

Texts that are not necessarily directly relevant to political science research but are used for context / complement. E.g. annotate etc.

- Wikipedia or other "ground truth" sources
- Network data
- Dictionaries: e.g. sentiment or emotions to use automated dictionary methods with one click

3.8 US Political Science focus

Item	Country	Description	Access	Link
General Social Survey	US	General Social Survey (GSS) monitors societal change in the US	Download: for SPSS, STATA	Here
The Supreme Court Database	US	Case Centered Data - Total Rows : 13,533	Download: CSV, DTA (STATA), POR (SPSS), RDATA, XLSX	Here
The Supreme Court Database	US	Justice Centered Data - Total Rows : 121,224	Download: CSV, DTA (STATA), POR (SPSS), RDATA, XLSX	Here
Congressional speech data	US	Congressional-speech corpus includes labels for whether the speaker supported or opposed, by-name references between speakers, and the scores that our agreement/disagreement classifier(s), debate and related extracted information. (9.8 Mb, tar.gz format)	Download: compressed tar.gz, multiple types including CSV	Here
ANES	US	Electoral behavior, political participation, and public opinion studies - Time Series Studies , Pilot Studies, Special Studies	Download	Here
CorPS	US	CORPS is a corpus of political speeches tagged with specific audience reactions, such as APPLAUSE or LAUGHTER.	Request from marco.guerini[at]trentorise.eu and strappa[at]fbk.eu	Here
Congressional Record for the 43rd-114th Congresses	US	Parsed Speeches and Phrase Counts	Download: zip of organized txt files	Here
GDELT	US	All events from broadcast, print, and web news from nearly every corner of every country in over 100 languages	Download: CSV	Here
The American Presidency Project	US	Presidential documents, papers, press, orders, memoranda etc	HTML	Here
Full text corpus data	US	10 large corpora of English: iWeb, COCA, COHA, NOW, Coronavirus, GloWbE, TV Corpus, Movie Corpus, Soap Corpus, Wikipedia	Purchase raw data in 3 formats	Here
GovInfo	US	Congressional Bills; Bill Status; Bill Summaries; Commerce Business Daily; Code of Federal Regulations (Annual Edition); Electronic Code of Federal Regulations; Federal Register; United States Government Manual; House Rules and Manual; Privacy Act Issuances; Public Papers of the Presidents of the United States; Supreme Court Decisions 1937-1975 (FLITE)	Download: XML	Here
DIME PLUS	US	Database on Ideology, Money in Politics, and Elections: Public version 2.0	Download: compressed CSV	Here
Replication data for: Tracing the	US	Replication Data	Download: com	Here

CHAPTER

FOUR

RESOURCES

This page contains links to resources, tools, courses, blogs, newsletters and other interesting things relating to text mining for research.

4.1 Tools and software

- This universal sentence encoder apparently is good for clustering sentences or shorter text
- A list of open source data labeling tools

See full list of 80+ tools (web apps, software, packages in R and python) with an overview of type and classification of free vs charged.

4.2 Books and Textbooks

- Computational Social Science with Python by Damian Trilling
- Language processing with Python , by Steven Bird, Ewan Klein and Edward Loper, Copyright © 2019

4.2.1 Course materials

- Big data and automated content analysis by Damian Trilling
- The python tutorial
- NLP for developers by Rasa
- Start to end training on wikipedia corpus for topic modeling
- Tips for computational text analysis from Simon Brown at Berkeley
- Lessons and materials for teaching text analysis from the Programming Historian
- An introduction on how to use pre-trained vector embeddings
- For working with shorter texts, like one sentences, where possibly there is just one topic and LDA assumes multiple topics, there is this GSDMM model with the full pipeline explained here
- Slide deck on Computational Analysis of Political Texts from the Data and Web Science Group at the Universit of Mannheim

4.3 Newsletters

- Sebastian Ruder's NLP News, probably the most comprehensive newsletter out there, covering deep dives in the technical and the economics of mining.
- The Gradient are overviews, essays and perspectives on Artificial Intelligence, recent developments and long-term impacts. It is a publication ran by volunteers and open to submissions.

4.4 Other interesting resources

- An overview of text mining in the social sciences and humanities by Dong Nguyen, arxiv preprint
- A critique of computational text analysis in the humanities
- An academic paper from the Workshop on Computational Humanities Research 2020, discussing the history of quantitative and computational research in the humanities, and especially the quantitative methods in history before computers; by Michael Piotrowski and Mateusz Fafinsky
- Estimating the degree of similarity between two texts, a blog by Adrien Sieg 2018
- Masakhane is a grassroots NLP community for Africans, by Africans. It brings people together to work on challenging research problems for African languages. Their recent EMNLP 2020 findings paper demonstrates the impact grassroots efforts can have.
- A super long and comprehensive list of great NLP resources by Keon

CHAPTER

FIVE

GENERAL

Note: SAGE Texti is still in beta, and may be updated frequently. Please use with caution, and let us know if you find any bugs.

Text currently works for pdf files, but doesn't support OCR.

If you want to report a **bug** or suggest an **enhancement**, please use this form . You will need to register with your google account to submit a screenshot. If you prefer not to, just email us directly, or raise an issue on GitHub. You can also contribute **new transformations** following these guidelines .

CHAPTER

SIX

TRY SAGE TEXTI

Go to SAGE Texti to create a new account, or get in touch with daniela (dot) duca (at) sagepub.co.uk.