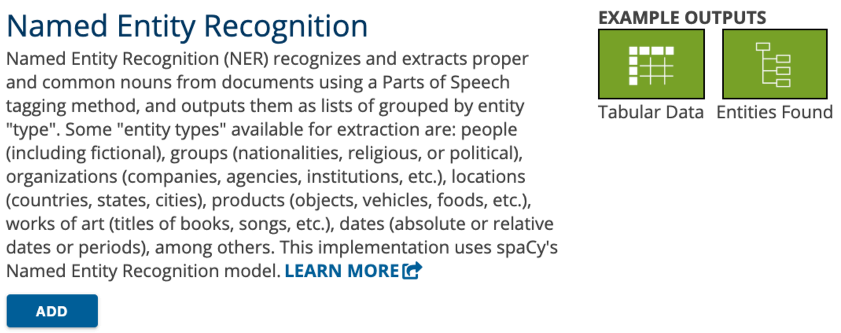
Digital Humanities in Practice

WEEK 8a Named Entity Recognition



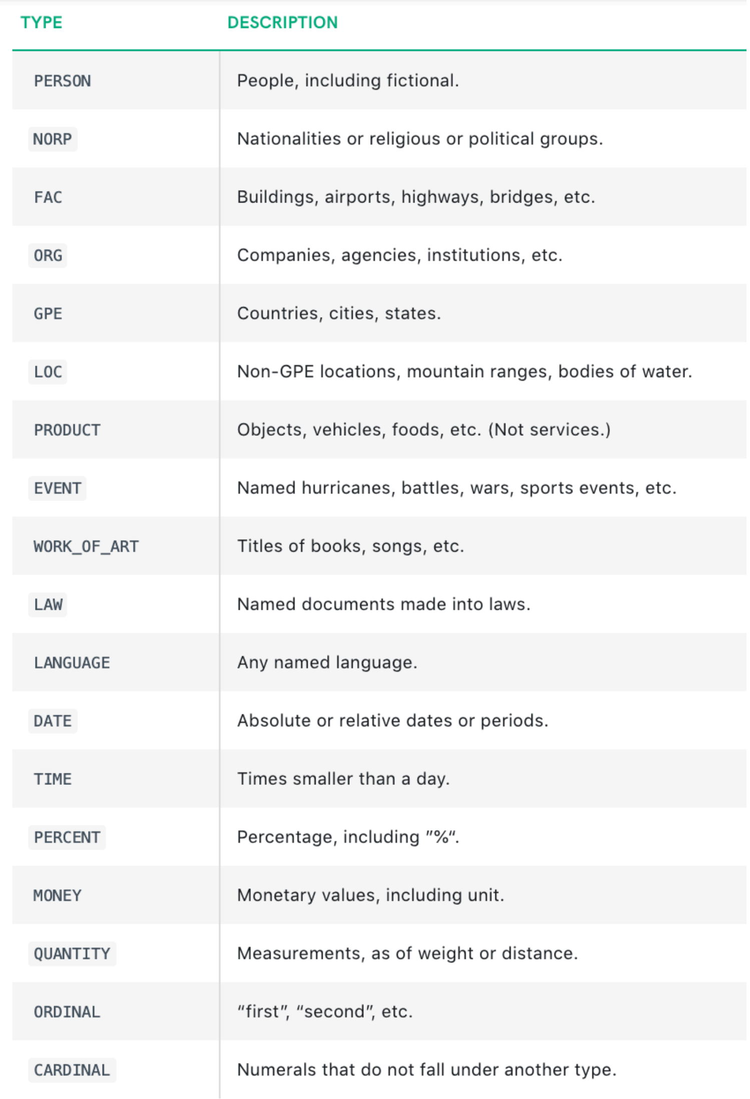
Named Entity Recognition

Named Entity Recognition (NER) recognizes and extracts proper and common nouns from documents using a Parts of Speech tagging method, and outputs them as lists of grouped by entity "type". Some "entity types" available for extraction are: people (including fictional), groups (nationalities, religious, or political), organizations (companies, agencies, institutions, etc.), locations (countries, states, cities), products (objects, vehicles, foods, etc.), works of art (titles of books, songs, etc.), dates (absolute or relative dates or periods), among others. This implementation uses spaCy's Named Entity Recognition model. **LEARN MORE**

Named Entity Recognition is often used to identify key people, places, and things within a Content Set. This tool can be useful when collecting data around place names for mapping, which can often be challenging to aggregate without the close reading of each document.

Digital Scholar Lab Implementation

The Named Entity Recognition tool is based on the open source spaCy model. When you run the tool, it will parse (work through) your content set, and identify words classified as 'entities' by the model, which as been trained using the OntoNotes 5 corpus. For your purposes, here is the list of entities which will be recognized, along with their abbreviation:



TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ”%“.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

Reading

Suvro Banerjee, 'Introduction to Named Entity Recognition', *Medium*, 2018 https://medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1

Kimmo Kettunen et al, 'Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910', *DHQ* 11:3, 2017 http://digitalhumanities.org/dhq/vol/11/3/000333/000333.html

Rainer Simon, Leif Isaksen, Elton Barker, and Pau de Soto Canamares, The Pleiades Gazetteer and the Pelagios Project' in Ruth Mostern, Humphrey Southall, Lex Berman, & Peter Bol, Placing Names: Enriching and Integrating Gazetteers, 2016. Project MUSE., https://muse.jhu.edu/.

Example Projects using Named Entity Recognition

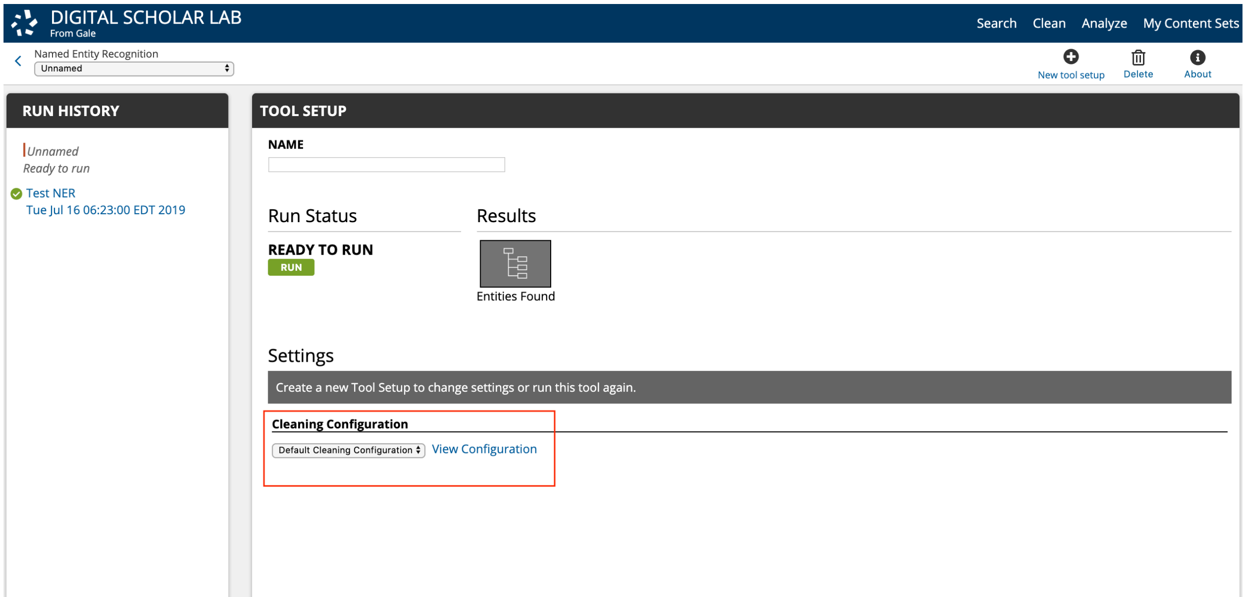
Gazetteers are one example of Named Entity Recognition in practice, in this case targeting place names in text. Often, this process is used as the starting point for building maps or historical narratives on the theme of 'place'.

Pelagios is one such project, which is well-funded and is very active. You can read more about it in the article, above. Check out the markup tool here: https://recogito.pelagios.org/

My own research is based on recognizing people and place names in Nile travel journals, letters and other ephemera from the late nineteenth century. My goal is to map social and travel networks in Egypt during this period which is known as the 'Golden Age of Egyptology'. This past year, I have been working with a Masters student in Computational Linguistics who has built a historical markup tool for our project, and others who want to use it. It essentially takes plain text (.txt) input, and then outputs text that has been encoded in a machine-readable format, in this case XML-TEI. The tool also identifies and marks up named entities in the text. You can try it out at our project website: http://www.emmabandrews.org/project/historical-markup-tool

Configuration Options in the Digital Scholar Lab

The only configuration option currently available for named entity recognition is the cleaning configuration you choose to apply to the content set.



Named Entity Recognition
Unnamed

Unnamed
Ready to run

Test NER
Tue Jul 16 06:23:00 EDT 2019

NAME

Run Status

READY TO RUN

Results

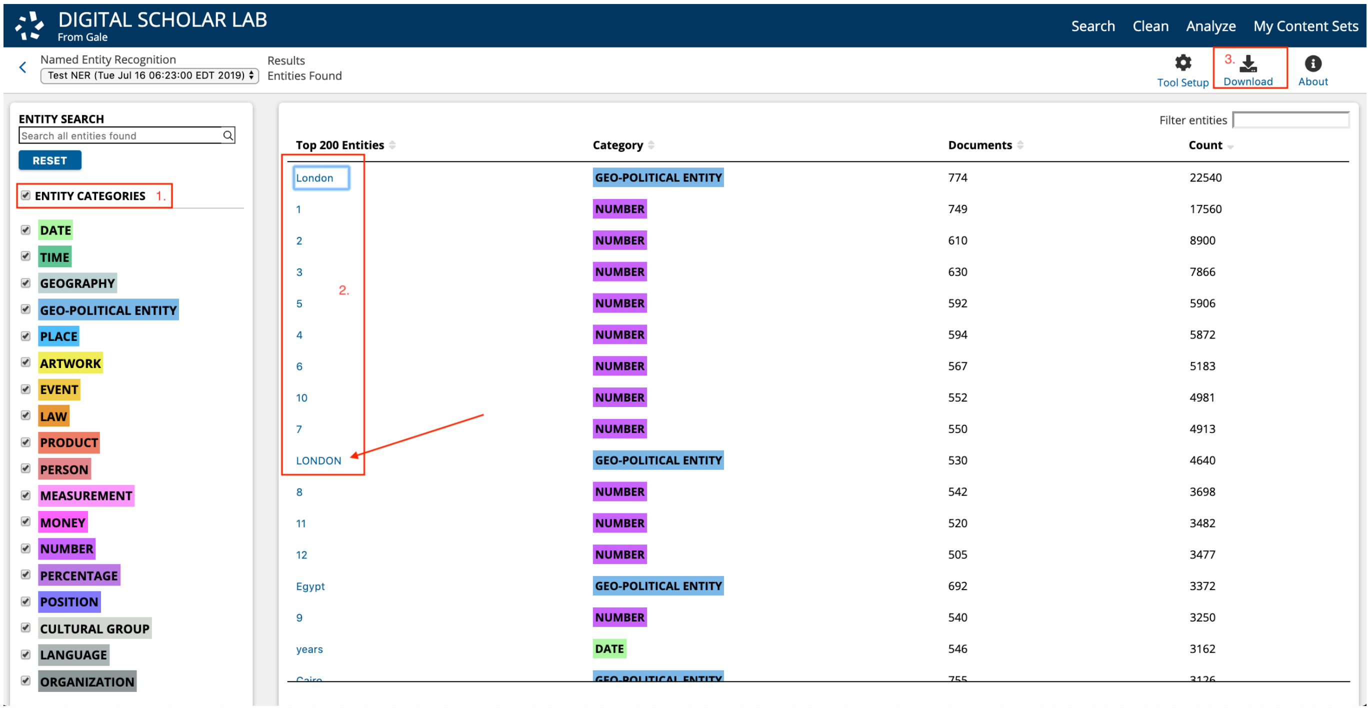
Settings

Create a new Tool Setup to change settings or run this tool again.

Cleaning Configuration

Default Cleaning Configuration View Configuration

Output and Visualizations



Named Entity Recognition
Test NER (Tue Jul 16 06:23:00 EDT 2019)
Results
Entities Found

☑ ENTITY CATEGORIES

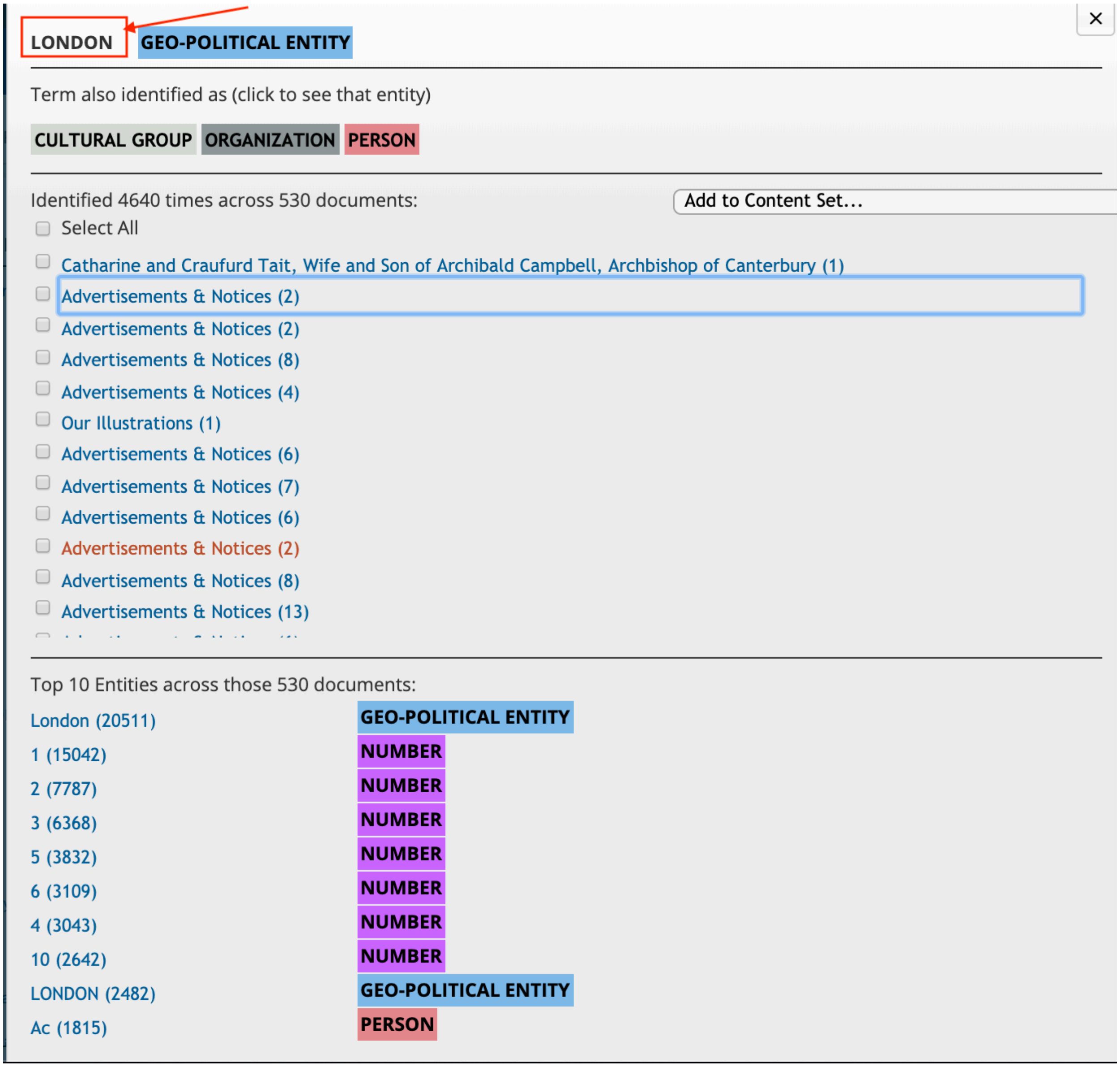
- ☑ DATE
- ☑ TIME
- ☑ GEOGRAPHY
- ☑ GEO-POLITICAL ENTITY
- ☑ PLACE
- ☑ ARTWORK
- ☑ EVENT
- ☑ LAW
- ☑ PRODUCT
- ☑ PERSON
- ☑ MEASUREMENT
- ☑ MONEY
- ☑ NUMBER
- ☑ PERCENTAGE
- ☑ POSITION
- ☑ CULTURAL GROUP
- ☑ LANGUAGE
- ☑ ORGANIZATION

Top 200 Entities	Category	Documents	Count
London	GEO-POLITICAL ENTITY	774	22540
1	NUMBER	749	17560
2	NUMBER	610	8900
3	NUMBER	630	7866
5	NUMBER	592	5906
4	NUMBER	594	5872
6	NUMBER	567	5183
10	NUMBER	552	4981
7	NUMBER	550	4913
LONDON	GEO-POLITICAL ENTITY	530	4640
8	NUMBER	542	3698
11	NUMBER	520	3482
12	NUMBER	505	3477
Egypt	GEO-POLITICAL ENTITY	692	3372
9	NUMBER	540	3250
years	DATE	546	3162
Cairo	GEO-POLITICAL ENTITY	755	3126

Each red box (above) is numbered:

1. In the left column, you'll see a list of Entity Categories identified by the tool in your content set, which are color coded. You can toggle each Entity Category on and off by checking or unchecking the boxes.

2. Individual entities are listed, along with the color-coded Entity Category, the number of documents the entity appears in, and the number of times it appears in total. The entities are clickable. Here, I've clicked into 'London':



LONDON GEO-POLITICAL ENTITY

×

Term also identified as (click to see that entity)

CULTURAL GROUP ORGANIZATION PERSON

Identified 4640 times across 530 documents:

Add to Content Set…

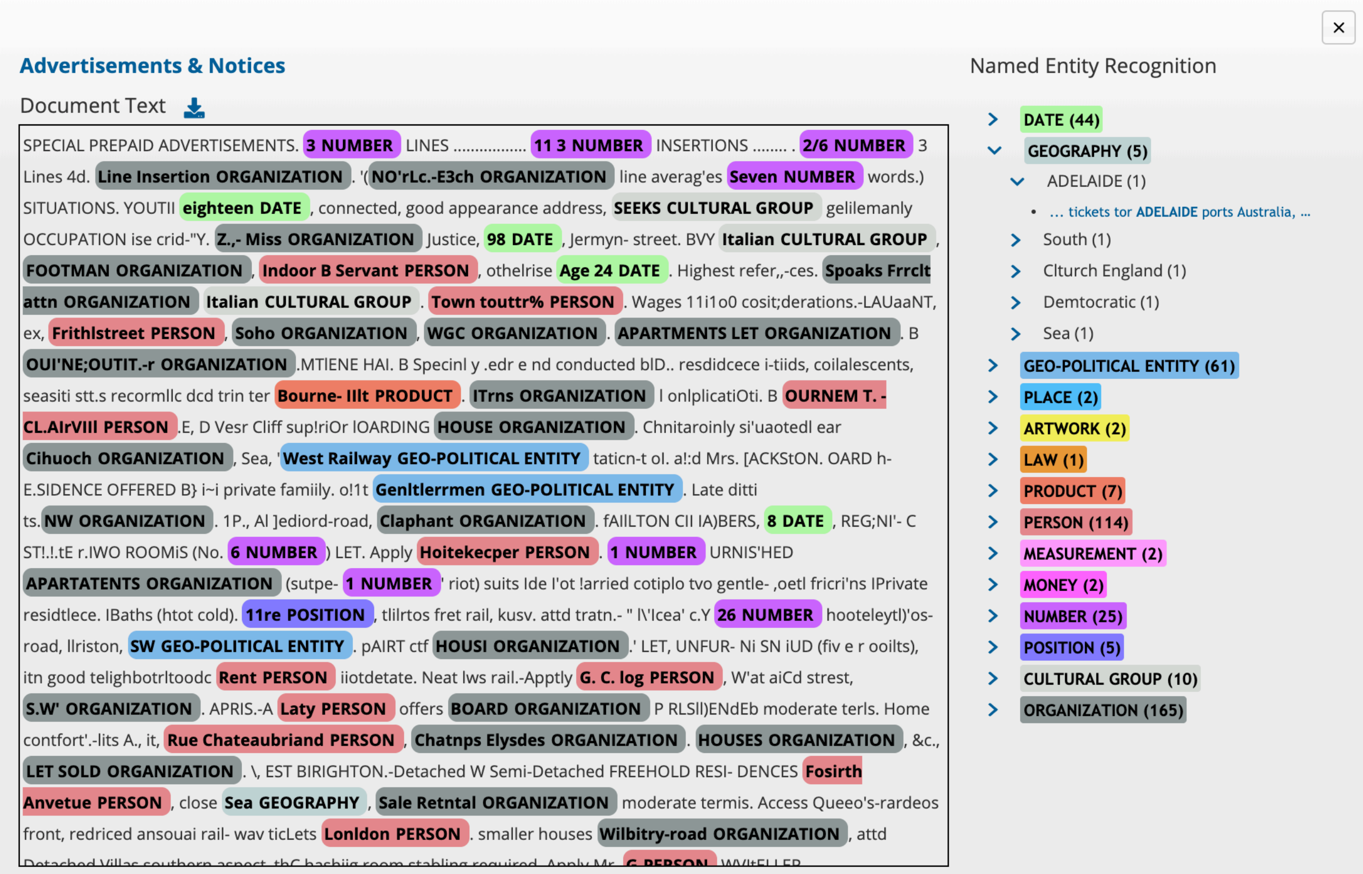
□ Select All

□ Catharine and Craufurd Tait, Wife and Son of Archibald Campbell, Archbishop of Canterbury (1)
□ Advertisements & Notices (2)
□ Advertisements & Notices (2)
□ Advertisements & Notices (8)
□ Advertisements & Notices (4)
□ Our Illustrations (1)
□ Advertisements & Notices (6)
□ Advertisements & Notices (7)
□ Advertisements & Notices (6)
□ Advertisements & Notices (2)
□ Advertisements & Notices (8)
□ Advertisements & Notices (13)

Top 10 Entities across those 530 documents:

London (20511)	GEO-POLITICAL ENTITY
1 (15042)	NUMBER
2 (7787)	NUMBER
3 (6368)	NUMBER
5 (3832)	NUMBER
6 (3109)	NUMBER
4 (3043)	NUMBER
10 (2642)	NUMBER
LONDON (2482)	GEO-POLITICAL ENTITY
Ac (1815)	PERSON

You can then click into individual documents in the list to look closely at what has been captured by the tool.



×

Advertisements & Notices

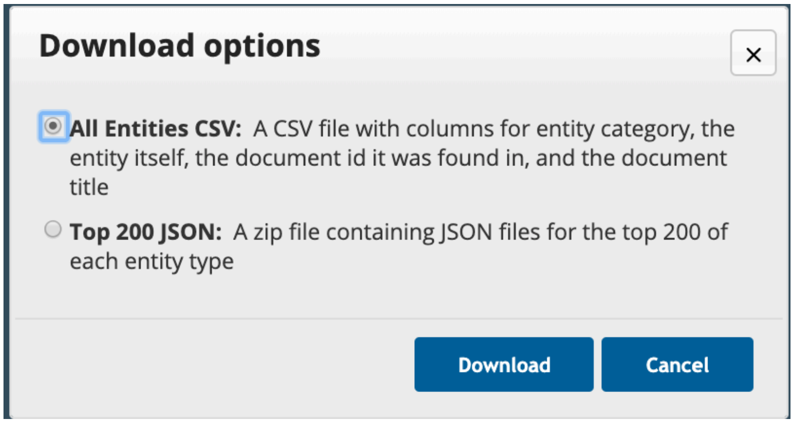
Document Text

SPECIAL PREPAID ADVERTISEMENTS. **3 NUMBER** LINES .... **11 3 NUMBER** INSERTIONS .... **2/6 NUMBER** 3 Lines 4d. **Line Insertion ORGANIZATION** . '(**NO'rLc.-E3ch ORGANIZATION** line averag'es **Seven NUMBER** words.) SITUATIONS. YOUTII **eighteen DATE** , connected, good appearance address, **SEEKS CULTURAL GROUP** gelilemanly OCCUPATION ise crid-"Y. **Z.,- Miss ORGANIZATION** Justice, **98 DATE** , Jermyn- street. BVY **Italian CULTURAL GROUP** , **FOOTMAN ORGANIZATION** , **Indoor B Servant PERSON** , othelrise **Age 24 DATE** . Highest refer,,-ces. **Spoaks Frrclt attn ORGANIZATION** **Italian CULTURAL GROUP** . **Town touttr% PERSON** . Wages 11i1o0 cosit;derations.-LAUaaNT, ex, **Frithlstreet PERSON** , **Soho ORGANIZATION** , **WGC ORGANIZATION** . **APARTMENTS LET ORGANIZATION** . B **OUI'NE;OUTIT.-r ORGANIZATION** .MTIENE HAI. B Specinl y .edr e nd conducted bID.. resdidcece i-tiids, coilalescents, seasiti stt.s recormllc dcd trin ter **Bourne- IIlt PRODUCT** . **ITrns ORGANIZATION** l onlplicatiOti. B **OURNEM T. - CL.AIrVIII PERSON** .E, D Vesr Cliff sup!riOr lOARDING **HOUSE ORGANIZATION** . Chnitaroinly si'uaotedl ear **Cihuoch ORGANIZATION** , Sea, '**West Railway GEO-POLITICAL ENTITY** taticn-t ol. a!:d Mrs. [ACKStON. OARD h-E.SIDENCE OFFERED B} i~i private famiily. o!1t **Genltlerrmen GEO-POLITICAL ENTITY** . Late ditti ts.**NW ORGANIZATION** . 1P., Al]ediord-road, **Claphant ORGANIZATION** . fAIILTON CII IA)BERS, **8 DATE** , REG;NI'- C ST!.!.tE r.IWO ROOMiS (No. **6 NUMBER**) LET. Apply **Hoitekecper PERSON** . **1 NUMBER** URNIS'HED **APARTATENTS ORGANIZATION** (sutpe- **1 NUMBER** ' riot) suits Ide l'ot !arried cotiplo tvo gentle- ,oetl fricri'ns IPrivate residtlece. IBaths (htot cold). **11re POSITION** , tlilrtos fret rail, kusv. attd tratn.- " l\'Icea' c.Y **26 NUMBER** hooteleytl)'os-road, llriston, **SW GEO-POLITICAL ENTITY** . pAIRT ctf **HOUSI ORGANIZATION** .' LET, UNFUR- Ni SN iUD (fiv e r ooilts), itn good telighbotrltoodc **Rent PERSON** iiotdetate. Neat lws rail.-Apptly **G. C. log PERSON** , W'at aiCd strest, **S.W' ORGANIZATION** . APRIS.-A **Laty PERSON** offers **BOARD ORGANIZATION** P RLSIl)ENdEb moderate terls. Home contfort'.-lits A., it, **Rue Chateaubriand PERSON** , **Chatnps Elysdes ORGANIZATION** . **HOUSES ORGANIZATION** , &c., **LET SOLD ORGANIZATION** . \, EST BIRIGHTON.-Detached W Semi-Detached FREEHOLD RESI- DENCES **Fosirth Anvetue PERSON** , close **Sea GEOGRAPHY** , **Sale Retntal ORGANIZATION** moderate termis. Access Queeo's-rardeos front, redriced ansouai rail- wav ticLets **Lonldon PERSON** . smaller houses **Wilbitry-road ORGANIZATION** , attd Detached Villas southern aspect, thC bashiig room stabling required. Apply Mr. **G PERSON** WWItELLER

Named Entity Recognition

- DATE (44)
- GEOGRAPHY (5)
  - ADELAIDE (1)
    - ... tickets tor **ADELAIDE** ports Australia, ...
  - South (1)
  - Clturch England (1)
  - Demtocratic (1)
  - Sea (1)
- GEO-POLITICAL ENTITY (61)
- PLACE (2)
- ARTWORK (2)
- LAW (1)
- PRODUCT (7)
- PERSON (114)
- MEASUREMENT (2)
- MONEY (2)
- NUMBER (25)
- POSITION (5)
- CULTURAL GROUP (10)
- ORGANIZATION (165)

3. You can download the full list in CSV or JSON format by clicking on the button.



Download options

×

● **All Entities CSV:** A CSV file with columns for entity category, the entity itself, the document id it was found in, and the document title

○ **Top 200 JSON:** A zip file containing JSON files for the top 200 of each entity type

Note: You will probably notice that some entities are misclassified by the tool. It's currently not possible to reclassify them in the DSL. The only way to do this is to download the CSV and then edit it.

Future updates of the NER tool will include the ability to edit and reclassified wrongly identified words.

Mapping

A natural companion to Named Entity Recognition is the mapping of place names, since they are readily extracted from text by running the analysis tool. Once the run is completed, a spreadsheet of Named Entities can be downloaded in the form of a CSV file, opened in Excel or OpenRefine and cleaned up to remove extraneous data. It's also possible to extract locations outside the DSL by exporting the OCR text and using the Clavin method, described below.

Once location data has been extracted, a number of tools exist for visualizing this data on a map. A few options are suggested below.

- Neatline - a plugin for Omeka. Tutorial here.
- StorymapJS - a free tool for creating narratives on a map base layer.
- Geolocator - activated in Omeka. Tutorial here
- Carto
- ArcGIS online
- ArcGIS Storymaps

Geoparsing Text Data

Working with plain text files, an option to extract location names is to use the online version of the CLAVIN geoparser (http://clavin.berico.us clavin-web/) to extract location data from select pages of a dataset.

CLAVIN (Cartographic Location And Vicinity INdexer) is an open source software tool for document geo-tagging and geo-parsing. It automatically extracts location names from structured and unstructured text and resolves them against a gazetteer to produce data-rich geographic entities. It has 75% accuracy for geospatial entity resolution, can resolve 100 locations per second per CPU and process 1 million documents in under an hour on a 9-node Hadoop cluster. It can scale to billions of records.

Full books may take a little while to generate geoparsed data - be patient. Caveat: you can only extract the top 20 locations (although on testing this, I note that I was able to extract 60!).

The goal is to submit text to the online geoparser and structure that information into an online spreadsheet, so that the results can be visualized using a mapping application.

The process

Step 1: Get the data

- Download your content set from the DSL. You have the ability to clean within the platform, which I recommend doing.

- You can either concatenate your individual .txt files into a single file, or upload several files one-by-one. Python file for concatenation is available here.

Step 2: Open the geoparser and the shared spreadsheet in web browser tabs

- Navigate your browser to the CLAVIN Web interface (http://clavin.berico.us/clavin-web/) *Note that depending on screen size, you may need to 'zoom out' your browser (Use the toolbar or Ctrl+scroll down), or resize the screen in order to move the map from covering the data results list.

- On a second tab, open the geolocation spreadsheet. **MAKE A COPY** in your own Drive.

*Option 1 Repeat steps 3-4 for each individual text file

*Option 2 Run your concatenated file through the geoparser, although this may take a while and will only return the top 20 results.

Step 3: Use the geoparser to extract location data

- Open the text file with a text editor

- Select all text (Ctrl+a on PC/Linux, for example) and copy it (Ctrl + c)

- Paste (Ctrl+v) the text into the text box on the online CLAVIN geoparser and click the 'Resolve Locations' button

Step 4: Copy and paste the results to your copy of the shared spreadsheet

- Highlight and copy the locations resolved by the geoparser
  - Do not copy the column headers ("ID|Name|Lat,Lon|Country Code|#")
- Navigate to YOUR COPY of the shared spreadsheet above and paste your results into open rows
  - Ensure that your data lines up with the headings

Download the spreadsheet to a working location on your computer.

Useful Tools for Mapping

Mapwarper: Find maps and other imagery, upload, and rectify against a real map.

Reed College Geocoding Application: https://rich.shinyapps.io/geocoder/

David Rumsey Historical Map Collection: https://amica.davidrumsey.com/home

NYPL Open Source Maps: http://www.openculture.com/2014/03/new-york-public-library-puts-20000-hi-res-maps-online.html

Chris Gist, 'Projection Lessons in Maps', *Scholar's Lab Blog*, December 1, 2011 https://scholarslab.lib.virginia.edu/blog/projection-lessons-in-maps/

Living Maps Review: http://livingmaps.review/journal/index.php/LMR/index

British Library Maps: https://www.bl.uk/maps/

Pelagios Network: https://pelagios.org/

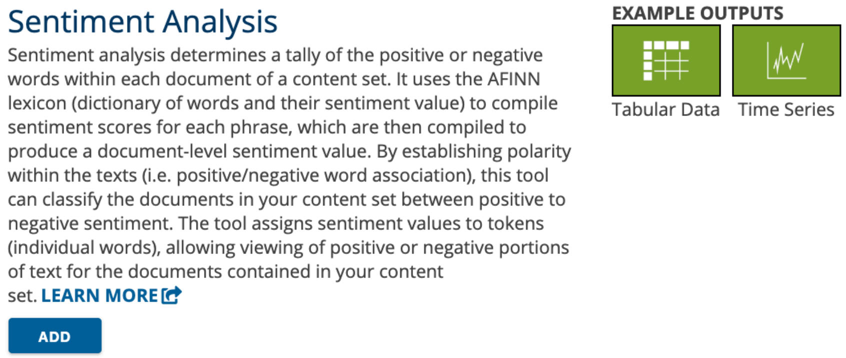
The Pelagios Network is a long-running initiative that links information online through common references to places. To create and maintain these connections, Pelagios has developed:

- *a method for creating semantic annotations, based on the W3C Web Annotation standard;*
- *tools and specifications for creating and making use of these annotations, most notably Recogito, an open-source platform for geo-annotating texts, images and databases;*

- *a community of individuals and organizations working with geographic data in humanities disciplines (history, language and literary studies, archaeology, etc.), and cultural heritage (galleries, libraries, archives and museums).*

Sentiment Analysis

Tool Overview



Sentiment Analysis

Sentiment analysis determines a tally of the positive or negative words within each document of a content set. It uses the AFINN lexicon (dictionary of words and their sentiment value) to compile sentiment scores for each phrase, which are then compiled to produce a document-level sentiment value. By establishing polarity within the texts (i.e. positive/negative word association), this tool can classify the documents in your content set between positive to negative sentiment. The tool assigns sentiment values to tokens (individual words), allowing viewing of positive or negative portions of text for the documents contained in your content set. **LEARN MORE**

Digital Scholar Lab Implementation

The DSL measures sentiment across time, based on the timeframe covered by your content set. It measures positive and negative sentiment based on the AFINN lexicon of positive and negative vocabulary. Future releases of the DSL will include the ability to leverage other sentiment lexicons, as well as including your own weightings.

Reading

- Jockers, Matthew L. “A Novel Method for Detecting Plot.” June 5, 2014. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/

This short blog post by Matt Jockers is one in a series where he delves into specific examples of how to use text analysis and visualizations for literary research. By tracing sentiment in the 19th century novel, Jockers (accidentally) “discovered that the sentiment I was detecting and measuring in the fiction could be used as a highly accurate proxy for plot movement.” In the post he describes how he uses sentiment analysis for detecting plot in four disparate novels: James Joyce’s *The Portrait of the Artist as a Young Man*, *Picture of Dorian Gray* by Oscar Wilde, *The Da Vinci Code* by Dan Brown, and Cormac McCarthy’s *Blood Meridian*. In a follow-up post to the one linked above, he describes how he normalizes “the plot shapes in 40,000 novels in order to compare the shapes and discover what appear to be six archetypal plots!”

- https://programminghistorian.org/en/lessons/sentiment-analysis

- Finn Årup Nielsen, AFINN Sentiment Lexicon, 2011 http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
- Parul Pandey, ‘Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)’, https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
- Sentiment Analysis with Python NLTK Text Classification (Try it out!)
- Acerbi A, Lampos V, Garnett P, Bentley RA, ‘The Expression of Emotions in 20th Century Books’. PLoS ONE 8(3): e59030, (2013) https://doi.org/10.1371/journal.pone.0059030

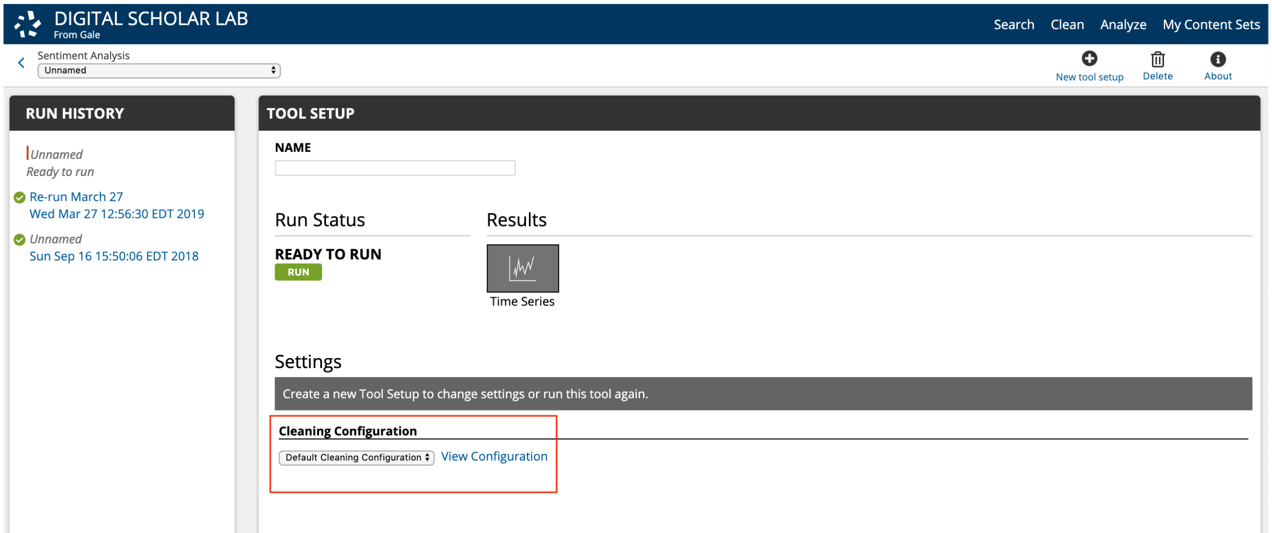
Example Projects Using Sentiment Analysis

Alexander Spangher, 'How does this article make you feel?', *New York Times* October 31, 2018 https://open.nytimes.com/how-does-this-article-make-you-feel-4684e5e9c47

'Sentiment analysis is opinion turned into code', *Open Objects* 2015 http://www.openobjects.org.uk/2015/04/sentiment-analysis-coming-to-a-newspaper-near-you/

Configuration Options in the DSL

The only configuration option currently available for sentiment analysis is the cleaning configuration you choose to apply to the content set.



Sentiment Analysis: Unnamed

Unnamed
Ready to run

Re-run March 27
Wed Mar 27 12:56:30 EDT 2019

Unnamed
Sun Sep 16 15:50:06 EDT 2018

NAME

Run Status

READY TO RUN

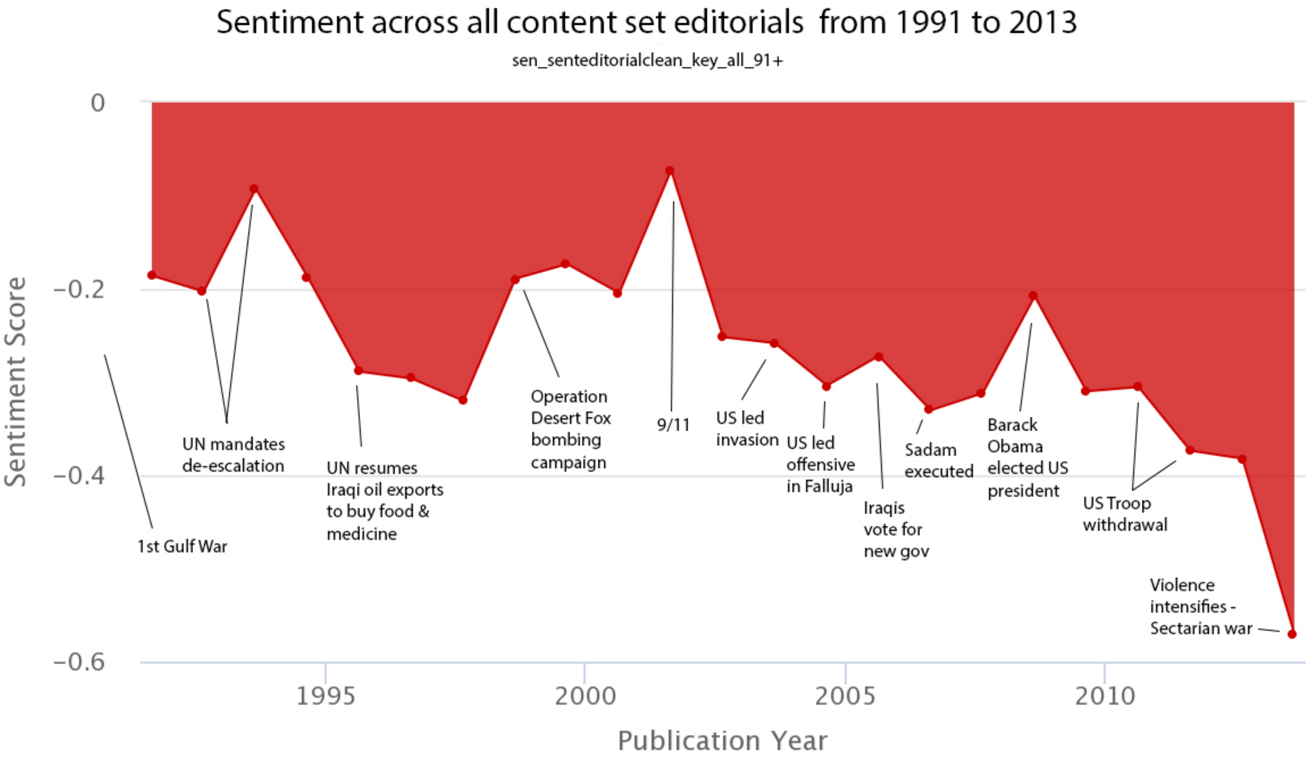
Results

Settings

Create a new Tool Setup to change settings or run this tool again.

Cleaning Configuration: Default Cleaning Configuration — View Configuration

When you run your analysis and generate the visualization, you'll be able to click into each point to take a closer look at the document to determine why it was considered positive or negative. Sometimes there are outliers which perhaps don't belong in your content set at all, and you can remove them at this stage of analysis.


Sentiment across all content set editorials from 1991 to 2013
sen_senteditorialclean_key_all_91+
Sentiment Score
Publication Year
1st Gulf War
UN mandates de-escalation
UN resumes Iraqi oil exports to buy food & medicine
Operation Desert Fox bombing campaign
9/11
US led invasion
US led offensive in Falluja
Iraqis vote for new gov
Sadam executed
Barack Obama elected US president
US Troop withdrawal
Violence intensifies - Sectarian war