

Natural Language Processing and Digital Humanities

Núria Bel

25-04-2018

<https://whatisdigitalhumanities.com/>

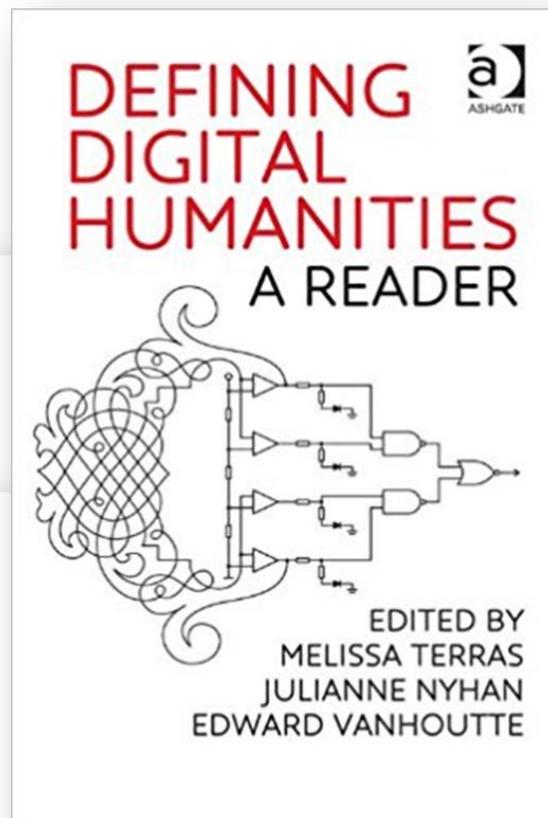
What Is Digital Humanities?

What Is Digital Humanities?

Doing research with digital materials and tools in a collaborative and open fashion. Not everyone needs to be a builder, though working with someone who can helps!

James Baker

NB: Refresh the page to get a new definition. Quotes were pulled from participants from the [Day of DH](#) between 2009-2014. As of January 2015, the database contains 817 rows and randomly selects a quote each time the page is loaded. If you want to do something cool with the data, I am providing a download for the CSV I compiled [here](#).



Italian Roberto Busa is considered the pioneer of Computational Linguistics. In 1946 he proposed a revolutionary idea to IBM: using computers to study texts, in particular the collected works of St. Thomas Aquinas. IBM decided to bet on the future.



New York - IBM World Headquarters: Presentation of the first result "Thomisticus" project

<http://www.corpusthomicum.org/it/ind>

Computers and the Humanities



Coverage: 1966-2004 (Vol. 1, No. 1 - Vol. 38, No. 4)
Published by: [Springer](#)

[Title History \(What is a title history?\)](#)

2005-2014 -

What has changed since 2004, when DH comes into the picture?

- Large efforts in digitalization that made many objects created by humans available for software tools.
- Other disciplines being hit by technologies that handle big data
- The irruption of Geographical Information Systems, GIS, and other visualization tools that have been used to easy 'interpretation'

Summary of DH research

Research in humanities is often related to the interpretation of objects created by humans

Such objects are mainly texts for literature and history, for instance

Research has to do with:

- Finding ... Texts
- Finding ... Data in texts
- Finding ... Patterns and relations in data, so that ..
- Making a particular interpretation in a sound and evident way is possible

Natural Language Processing (NLP) are available tools for ...

- Finding data in **large quantities of texts**
- Finding and **extracting patterns and correlations**
- Representing data to enable their analysis

NLP tools find names, relations, opinions ...

▼ Language identification

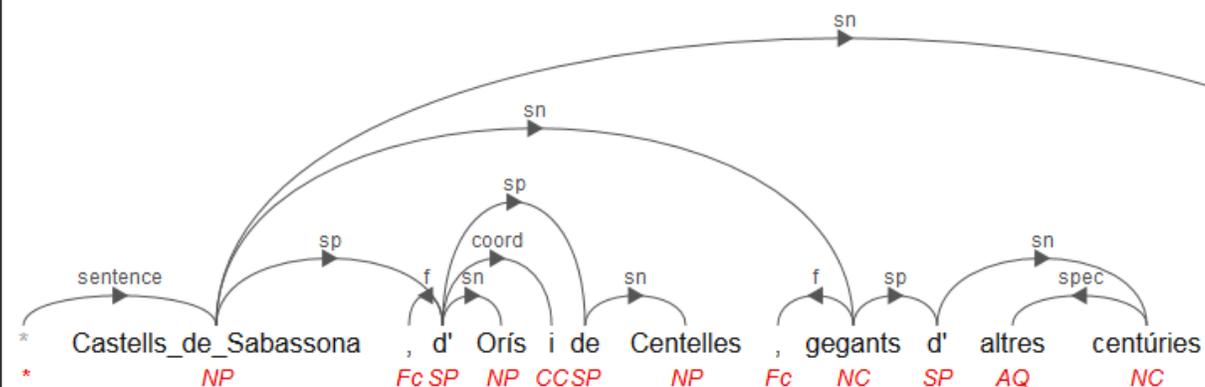
User selected language is: Catalan (ca)

FreeLing 4.0 - An Open-Source Suite of Language Analyzers
Hooked on a FreeLing?

▼ Sentences

Sentence 1

Castells_de_Sabassona	,	d'	Orís	i	de	Centelles	,	gegants	d'
castells_de_sabassona		de	orís	i	de	centelles		gegant	de
NP00G00		Fc SP	NP00G00	CC	SP	NP00G00		Fc	NCMP000 SP



Lluís Padró and Evgeny Stanilovsky.

FreeLing 3.0: Towards Wider Multilinguality

Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.



Detailed Sentiment Analysis

Chapter 1.

It was a bright cold day in April, and the clocks were striking thirteen.

Winston Smith, his chin nuzzled into his breast in an effort to escape the

vile wind, slipped quickly through the glass doors of Victory Mansions,.

though not quickly enough to prevent a swirl of gritty dust from entering

along with him.

The hallway smelt of boiled cabbage and old rag mats.

Moreno-Ortiz, A. (2016). Lingmotif 1.0 [Computer Software]. Málaga, Spain: Universidad de

Málaga. Retrieved from Tecnolengua website: <http://tecnolengua.uma.es/lingmotif>

Finding names. Examples

Geographical Information Systems (GIS) and NLP Named Entity Recognition & Desambiguation (NER)

The task is:

- identifying an entity mention (a proper noun) and
- identifying to which entity it refers to.

Geographical Names (assigning coordinates, for instance)
Person Names (linking to VIAF, Wikipedia, for instance)

	A	B	C
1	Place	Coordinates	
2	Amsterdam	52.373056, 4.892222	
3	Ankara	39.916667, 32.833333	
4	Arlington	38.880278, -77.108333	
5	Athens	37.966667, 23.716667	
6	Baden Baden	48.762778, 8.240833	
7	Belgrade	44.816667, 20.466667	
8	Berlin	52.516667, 13.383333	
9	Bern	46.95, 7.45	
10	Boston	42.358056, -71.063611	
11	Brussels	50.85, 4.35	
12	Bucharest	44.4325, 26.103889	
13	Budapest	47.471944, 19.050278	
14	Bristol	51.45, -2.583333	

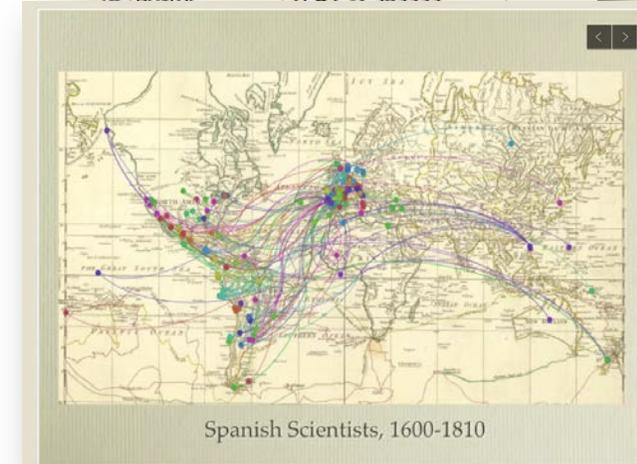
VIAF
Fichero de Autoridades Virtual Internacional

Búsqueda

Seleccione campo: Todos los encabezamientos
Seleccione índice: Todo VIAF
Términos de búsqueda:
Búsqueda

1,913 resultados encontrados para **Cervantes**

Encabezamiento	Tipo	Título de muestra
1 Cervantes Saavedra, Miguel de 1547-1616	Autor personal	Amante liberal Don Quixote



GIS and NER (1)

Pelagios Mediaeval Iberia <http://commons.pelagios.org>

Annotating texts with Identification of Named Entities

- NER with FreeLing
- Relate and extend with LOD, dbpedia...
- And geolocation with Geographic Information Systems (GIS)

The image shows a screenshot of the Pelagios Commons web application. On the left, there is a list of text files (e.g., KT.ACN.txt, KT.AAN.txt). The main area displays a text document with several named entities highlighted in green: Xerez, Sevilla, Toledo, and Heznalcazar. A modal window is open over the text, showing a map of Iberia with a location pin. The modal includes fields for 'Place', 'Person', and 'Event', and a search bar. The Pelagios Commons logo and navigation menu are visible on the right side of the interface. The navigation menu includes links for 'About', 'Link Data', 'Explore Data', 'Pelagios Commons', and 'Community Activity'. The footer of the page features the UPF logo and the text 'Universitat Pompeu Fabra Barcelona'.

GIS and NER (2)



ORGANIZACION DE LOS INDICADORES POR MUNICIPIO

AYUNTAMIENTOS.	CÉDULAS DE INSCRIPCIÓN.	HABITANTES.	POR NATURALEZA.				POR SEXO.	POR ESTADO CIVIL.							
			NACIONALES.		EXTRANJEROS.			Solteros.	Casados.	Viudos.	Menos de 1 año.	De 1 a 5.	De 6 a 10.	De 11 a 15.	De 16 a 19.
			Matriculados.	Transmigrados.	Matriculados.	Transmigrados.									
PARTIDO DE ARENS DE MAR.															
ARENS DE MAR.....	1.105	5.219	Varones . 2.328 Hembras . 2.827	28 29	2 ..	5 ..	2.363 2.876	1.363 1.595	877 963	123 298	70 72	297 267	276 258	218 257	133 232
ARENS DE MUNT.....	601	3.305	Varones . 1.642 Hembras . 1.624	24 15	1.666 1.639	1.003 907	616 619	47 113	50 55	908 903	199 160	174 188	122 152
CALPELLA.....	774	3.526	Varones . 1.585 Hembras . 1.882	38 19	1 1	1.624 1.902	915 1.030	611 684	68 188	48 37	183 170	158 196	161 190	106 183
CAMPINS.....	72	373	Varones . 209 Hembras . 160	4	213 160	134 80	73 70	6 10	5 7	25 12	27 13	16 15	13 8
CANET DE MAR.....	724	3.232	Varones . 1.518 Hembras . 1.694	1.538 1.694	937 1.006	544 564	57 124	48 42	161 170	190 205	138 130	124 116
FOGAS DE TORDERA.....	143	707	Varones . 319 Hembras . 292	71 18	4 3	394 313	247 169	129 125	18 19	12 9	38 45	47 38	41 28	26 20

ORIGINAL RESEARCH ARTICLE

Front. Digit. Humanit., 04 October 2017 | <https://doi.org/10.3389/fdigh.2017.00019>



Heritage As a Source of Studies into Industrial History: Using Digital Tools to Explore the Geography of the Industrialization

Guillermo Esteban-Oliver*, Adrià San José* and Jordi Martí-Henneberg*

Department of Geography and Sociology, University of Lleida, Lleida, Spain

The main objective of this article is to explore the possibility of combining two very different sources in order to study the distribution of industrial activity throughout history. The traditional primary sources to use for this purpose are the official censuses on population and economic activity that have been conducted in the majority of countries since the mid-nineteenth century. However, the majority of these lack detail at the regional level and also with respect to the types of professional occupations that they quantify. In order to complement and profile these census data, we propose the use of another type of information which can also be quantified, but whose characteristics are very different. We refer to the industrial heritage sites identified in digital format in a given territory, which in this case is Catalonia, Spain. This innovative dataset was obtained using digital tools such as web scraping and data mining techniques. This type of historical information was used to check whether it is reliable and valid for interpreting the spatial impact of the introduction of industrial activity. The article also shows that the systematic identification of elements of industrial heritage offers a new and very useful source of information for interpreting the history of industrial geography.

Materials and Methods

The methodology used to develop this research consisted of an analysis to correlate industrial occupations with data about industrial heritage sites. We will start by first describing the methodology that allowed us to compile a database for industrial occupations in 1860. Then we will describe how we created a complementary database on industrial sites. As a result, we were able to analyze whether the two datasets were correlated and the extent to which they were able to provide detailed georeferenced data.⁶

Finding patters. Examples

NLP for finding patterns

The corpus has more than 5,000 sonnets (about 71,000 lines)

ADSO *Distant reading approach to Golden Age Spanish Sonnets*

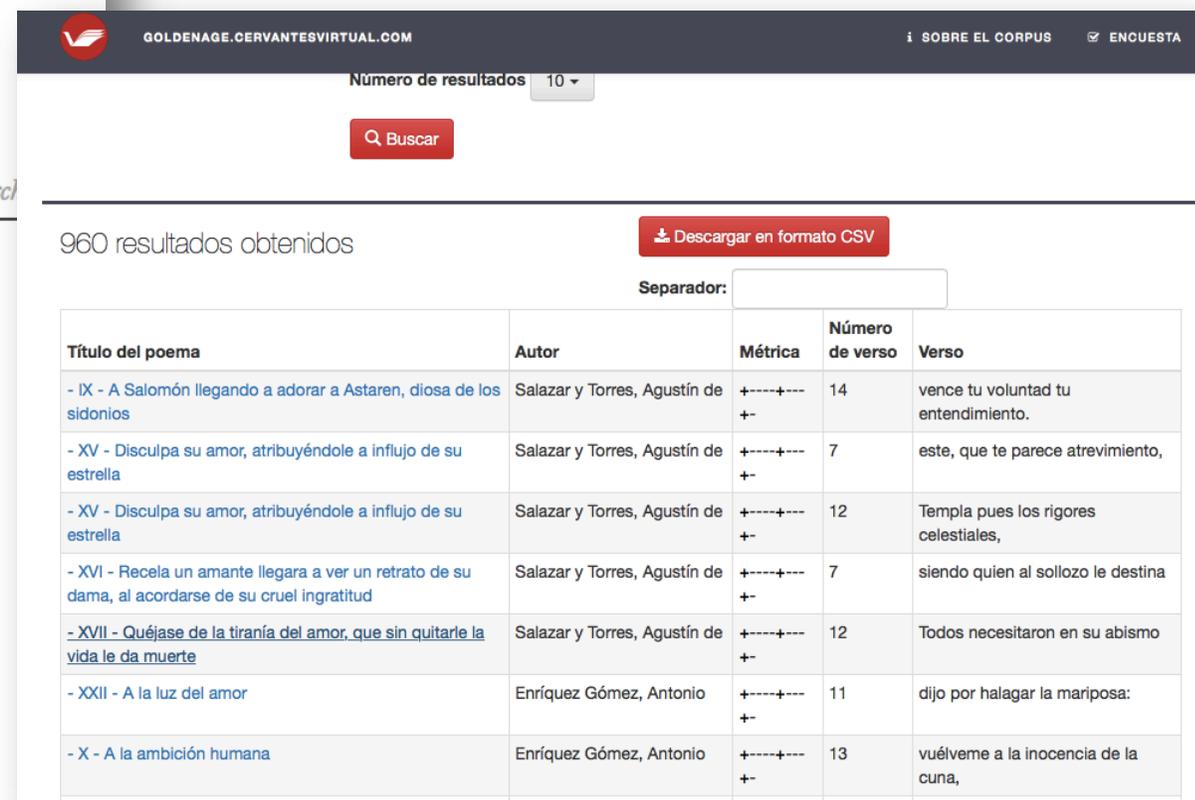
ADSO project ▾ Corpus of metre ▾

ADSO project

The purpose of this Project is to develop a macroanalysis and distant reading of Spanish Golden Age Sonnets, from the Renaissance Era (Garcilaso de la Vega) to the poetry from the end of the Baroque period (Sor Juana Inés de la Cruz). Computational methods will be used in order to perform the analysis, so that the main recurrent traits —both metrical and semantic— can be detected and singled out.

Unlike previous studies, our aim is not to analyze the limited number of sonnets that have been standardized as ‘canonical’ poetry, but rather to achieve the identification and characterization of those literary traits, both metrical and semantic, that all the sonnets from the 16th and 17th centuries share (Navarro Colorado, 2015, 2016), using computational techniques.

Search



GOLDENAGE.CERVANTESVIRTUAL.COM SOBRE EL CORPUS ENCUESTA

Número de resultados 10 ▾

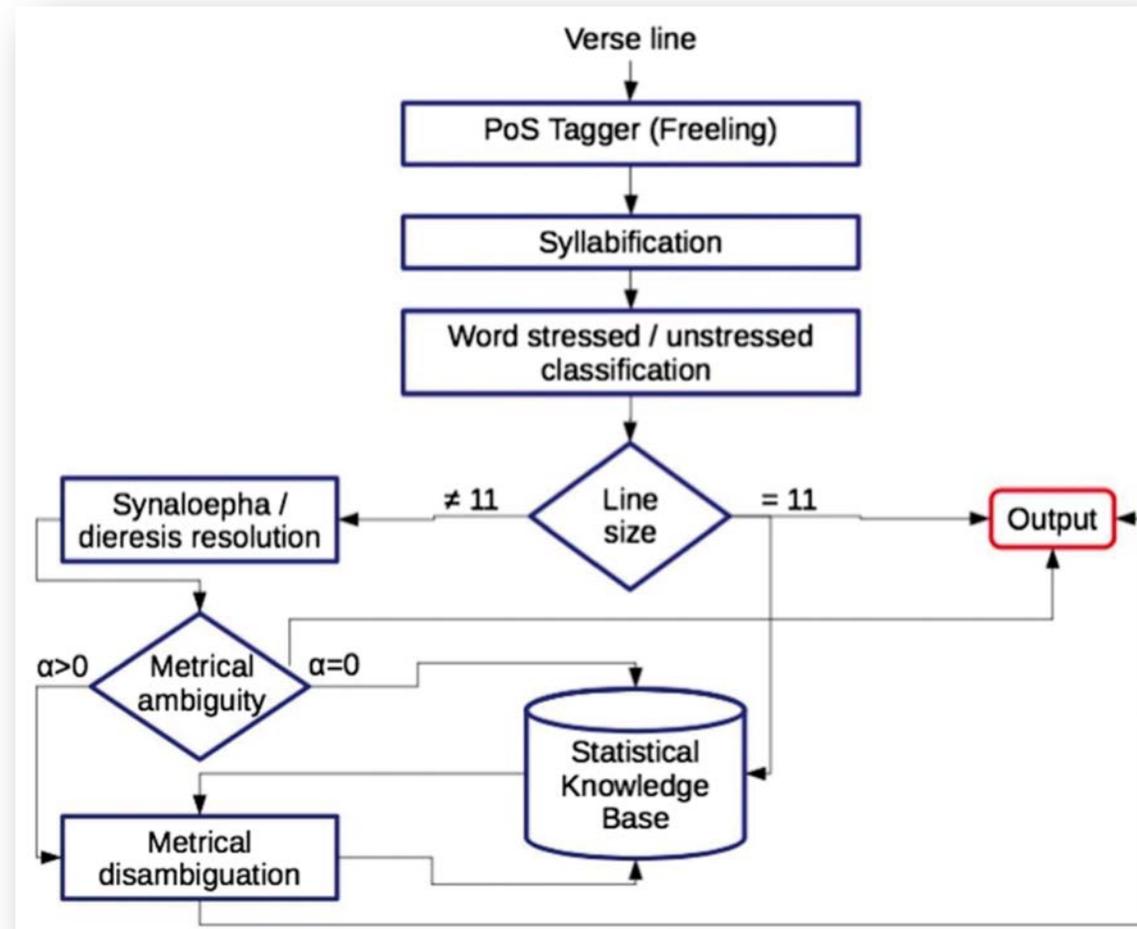
Buscar

960 resultados obtenidos Descargar en formato CSV

Separador:

Título del poema	Autor	Métrica	Número de verso	Verso
- IX - A Salomón llegando a adorar a Astaren, diosa de los sidonios	Salazar y Torres, Agustín de	+----+ +	14	vence tu voluntad tu entendimiento.
- XV - Disculpa su amor, atribuyéndole a influjo de su estrella	Salazar y Torres, Agustín de	+----+ +	7	este, que te parece atrevimiento,
- XV - Disculpa su amor, atribuyéndole a influjo de su estrella	Salazar y Torres, Agustín de	+----+ +	12	Templa pues los rigores celestiales,
- XVI - Recela un amante llegara a ver un retrato de su dama, al acordarse de su cruel ingratitude	Salazar y Torres, Agustín de	+----+ +	7	siendo quien al sollozo le destina
- XVII - Quéjase de la tiranía del amor, que sin quitarle la vida le da muerte	Salazar y Torres, Agustín de	+----+ +	12	Todos necesitaron en su abismo
- XXII - A la luz del amor	Enríquez Gómez, Antonio	+----+ +	11	dijo por halagar la mariposa:
- X - A la ambición humana	Enríquez Gómez, Antonio	+----+ +	13	vuélveme a la inocencia de la cuna,

Borja Navarro Colorado (2018) "A Metrical Scansion System for Fixed-Metre Spanish Poetry", *Digital Scholarship in the Humanities*, Volume 33 (1), pages 112–127



NLP for Sentiment Analysis. Example

NLP

Sentiment Analysis & Text Classification

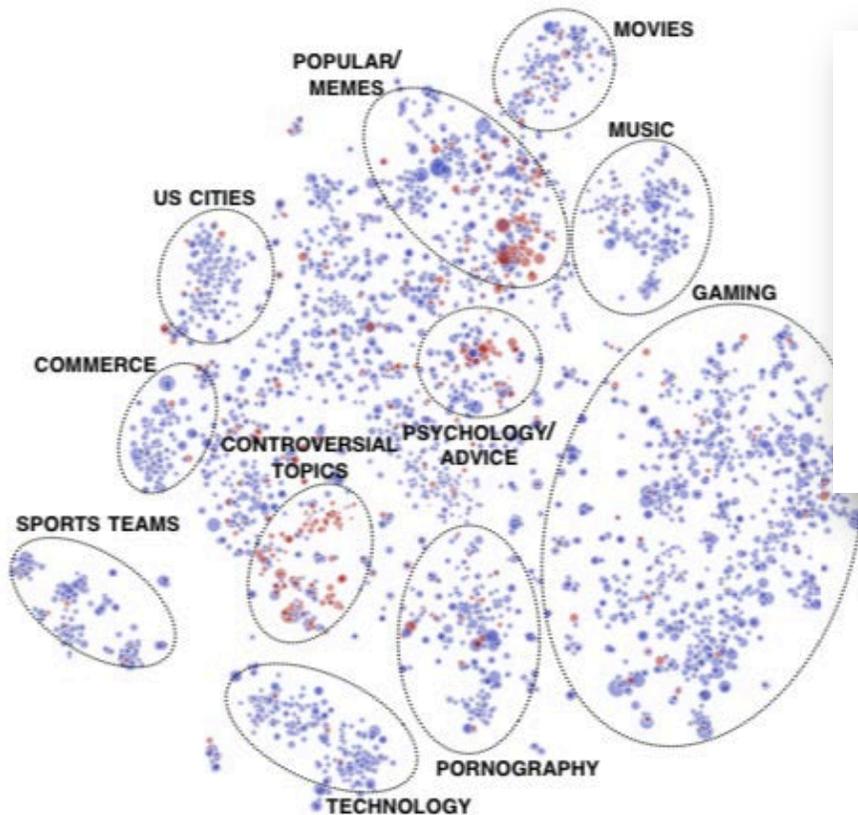


Figure 1: Communities in Reddit: each node represents a community. Red nodes initiate more conflicts, while blue nodes do not. Communities are embedded using user-community information, as described in Section 6. Figure best viewed in color.

Community Interaction and Conflict on the Web

Srijan Kumar
Stanford University, USA
srijan@cs.stanford.edu

Jure Leskovec
Stanford University, USA
jure@cs.stanford.edu

William L. Hamilton
Stanford University, USA
wleif@stanford.edu

Dan Jurafsky
Stanford University, USA
jurafsky@stanford.edu

It analyzed 40 months of data, containing 1.8 billion comments made by over 100 million users across 36,000 communities

Srijan Kumar, William L. Hamilton, Jure Leskovec, Dan Jurafsky. 2018. [Community Interaction and Conflict on the Web](#). Proceedings of the Web Conference (WWW). 2018 (to appear).

NLP for Data Analysis. Examples

NLP for data analysis



In screen direction from film scripts,

“she” is more likely to...



and “he” is more likely to...



The curious data here is less what Rose says (“I’m flying”) and more what the screen direction prescribes (“she smiles dreamily,” “he pushes against her”). In the following analysis, we go deep on screen direction to understand gender tropes. **We examined 2,000 scripts and broke down every screen direction mapped to the pronouns “she” and “he.”**

The R code behind this analysis is [publicly available on GitHub](#).

<https://juliasilge.com/about>

<https://pudding.cool/2017/08/screen-direction/>

Word embeddings quantify 100 years of gender and ethnic stereotypes

Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS 201720347 (2018).
doi:10.1073/pnas.1720347115

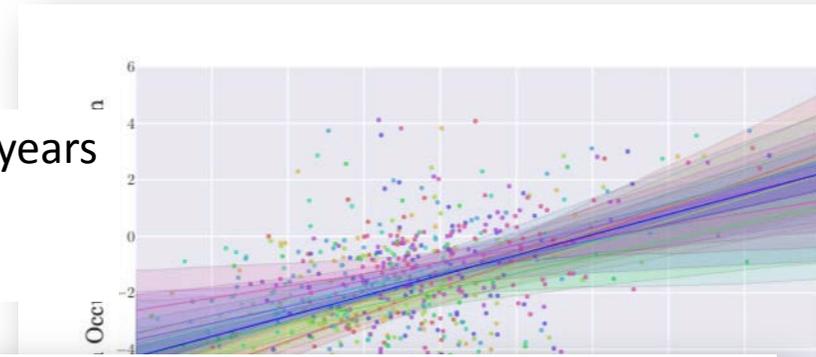


Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian
Housekeeper	Professor
Mason	Official
Artist	Secretary
Janitor	Conductor
Dancer	Physicist
Mechanic	Scientist
Photographer	Chemist
Baker	Tailor
Cashier	Accountant
Driver	Engineer

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

NLP and Digital Humanities?

It's impossible to conduct research without software, say 7 out of 10 UK researchers

The Software Sustainability Institute, University of Edinburgh

(<http://www.software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers>)

2014	Do you use research software?			Do you develop your own research software?		
	Yes	No	%	Yes	No	%
Humanities & Language Based Studies & archaeology	Yes	21	55,3%	Yes	5	13,2%
	No	17	44,7%	No	33	86,8%
Medicine, Dentistry & Health	Yes	55	96,5%	Yes	24	42,1%
	No	2	3,5%	No	33	57,9%
Social Studies	Yes	38	86,4%	Yes	15	34,1%
	No	6	13,6%	No	29	65,9%

Researchers need information about tools

The image is a collage of three screenshots from digital humanities tool websites. The top screenshot is from MIT Libraries, showing a navigation bar with 'MIT Libraries' and links for 'Search', 'Hours & locations', 'Borrow & request', 'Research support', and 'About us'. Below the navigation bar is a section titled 'Digital Humanities : Tools' with sub-sections for 'Home', 'DH Projects (MIT and more)', 'News and Community', 'Tools', and 'Training'. The 'Tools' sub-section is active, displaying a list of tool categories: 'Tool Boxes, Registries, Crawlers, & more!', 'Data Mining, Text Encoding, and Text Analysis Tools (in alphabetical order)', and 'Web Publishing and Timelines'. The 'Data Mining...' section lists 'Annotation Studio' and 'Juxta'. The 'Web Publishing...' section lists 'Omeka'. The middle screenshot is from Cambridge University Library, showing a navigation bar with '600 YEARS CAMBRIDGE UNIVERSITY LIBRARY 1416 - 2016' and links for 'Home', 'Visiting the Library', 'Research', 'What's On', 'Search & find', 'Collections', 'About', 'Giving', 'Cambridge Libraries', and 'Contact'. Below the navigation bar is a section titled 'Digital Humanities Tools' with a sidebar menu for 'Cambridge University Library', 'Research', 'Digital Humanities', and 'More Info'. The 'More Info' section is expanded, showing 'Tools', 'Resources', and 'List of projects'. The bottom screenshot is from UCLA Library, showing a navigation bar with 'UCLA Library' and links for 'Connect from Off-Campus', 'Hours', 'Contact', and 'Ask a Librarian'. Below the navigation bar is a section titled 'Digital Humanities' with a search bar and a sidebar menu for 'Home', 'Reference', 'Publications', 'Centers', 'Programs', 'Projects', 'Tools', 'Community', 'Workshops', and 'Digital Content'. The 'Tools' section is active, displaying a list of tool categories: 'Tools and Methods', 'Methods', 'Open Source Tools', 'Commercial Tools', and 'Research Workshops at UCLA'. The 'Tools and Methods' section is expanded, showing a list of tools: 'Bamboo DIRT', 'arts-humanities.net', and 'Lincoln Logarithms'.

MIT Libraries

Search Hours & locations Borrow & request Research support About us ASK US ACCOUNT

Digital Humanities : Tools

Home DH Projects (MIT and more) News and Community Tools Training

Tool Boxes, Registries, Crawlers, & more!

- Bamboo DiRT**
A tool, service, and collection registry of digital research tools for scholarly use.
- Bibliopedia**
Bibliopedia will perform advanced data-mining, cross-referencing, and scholarly literature centered collaboration.
- Color Brewer**
Find the right color for your maps and mapping tools.
- Digital Toy Chest for Humanists**

Data Mining, Text Encoding, and Text Analysis Tools (in alphabetical order)

- Annotation Studio**
A suite of collaborative web-based annotation tools.
- Juxta**
Juxta is an open-source tool for comparing and collating multiple witnesses to a single textual work. Originally designed to aid scholars and editors examine the history of a text from manuscript to print versions, Juxta offers a number of possibilities for

Web Publishing and Timelines

- Omeka**
Omeka is a free and open-source publishing platform that allows collaborators to display content and build digital timelines. Available for installation on your server or via [more...](#)
- Scalar**
Great for narrative support, novice and advanced users

UCLA Library

Connect from Off-Campus Hours Contact Ask a Librarian

Digital Humanities

UCLA Library » Research Guides » Digital Humanities » Tools

Search this Guide

Tools and Methods

In the Digital Humanities, methods like data analysis, data capture, and data structuring allow digital humanists to find patterns, search across large bodies of text, and engage in forms of scholarship that were not previously possible. Projects are an essential aspect of digital humanities for it is here that the interdisciplinary nature of Digital Humanities is most apparent. Digital Humanities projects reflect an intersection of academic disciplines and exist to answer humanities based questions by integrating a variety of multimedia formats in a dynamic environment. The tools and projects listed below exemplify these exciting new forms of humanities work.

Methods

The term "methods" refers to the computer based techniques used to create, analyze and disseminate digital resources. From animation and publishing tools to techniques for extracting and analyzing data, the methods employed by digital humanists are abundant and overlapping. To learn more, browse the "methods" of [arts-humanities.net](#).

- Bamboo DIRT**
Tools imply methods. Bamboo DIRT organizes tools by DH tasks.
- arts-humanities.net**
- Lincoln Logarithms**
a new project from the Digital Scholarship Commons at Emory University, uses four text analysis tools, MALLET, Voyant, Paper Machines, and Viewshare, to examine 57 full text sermons given on the occasion of Lincoln's assassination. Interesting enough in its own right, the project also explicitly addresses some of the major obstacles in DH projects.

CAMBRIDGE UNIVERSITY LIBRARY

600 YEARS 1416 - 2016

Home Visiting the Library Research What's On Search & find Collections About Giving Cambridge Libraries Contact

Digital Humanities Tools

Cambridge University Library

Research

Digital Humanities

More Info

- > Tools
- > Resources
- > List of projects

About

How can we help?

Related links

- Cambridge Digital Library
- Digital Content Unit
- Office of Scholarly Communication
- Cambridge Digital Humanities Network
- Contact us

UCLA Library

UCLA Library » Research Guides » Digital Humanities » Tools

Digital Humanities

Search this Guide

Tools

- Tools and Methods
- Methods
- Open Source Tools
- Commercial Tools
- Research Workshops at UCLA

Community

- Workshops
- Digital Content

About research infrastructures ...

<https://www.clarin.eu>



The screenshot shows the CLARIN website homepage. At the top, there is a navigation menu with links for 'About', 'Participants', 'Services', 'Knowledge Base', 'Events', 'News', and 'Contact'. The main heading is 'CLARIN - European Research Infrastructure for Language Resources and Technology'. Below this, a paragraph describes the infrastructure's mission: 'CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area. [Read more...](#)'. To the right of the text is the CLARIN logo, which consists of a stylized network of nodes and lines above the text 'CLARIN' and 'Common Language Resources and Technology Infrastructure'.



By Researchers for Researchers

DARIAH is a pan-european infrastructure for arts and humanities scholars working with computational methods. It supports digital research as well as the teaching of digital research methods.

<https://www.dariah.eu>

Summary of DH research and NLP

Research in humanities is often related to the interpretation of objects created by humans

Such objects are mainly texts for literature and history, for instance

Research has to do with:

- Finding ... Texts ????????
- ✓ Finding ... Data in texts
- ✓ Finding ... Patterns and relations in data, so that ..
- ✓ Making a particular interpretation, sound and evident, is possible

Finding texts: Researchers need downloadable and processable DATA as collections ...

The image shows a screenshot of the Biblioteca Digital Hispánica website. A modal window is open for selecting download formats. The modal is titled "Descargas" and asks the user to "Elija el formato del fichero:". It offers three main format options: PDF, JPEG, and TXT (text no estructurado). The "TXT" option is circled in red. Under each format, there are radio buttons for "de esta página" and "de varias páginas" (with a page range input field). Below the format selection, there is a note: "El texto se obtiene a partir de un proceso de OCR y su calidad puede variar en función de la tipografía original del documento." and a "Descargar" button. At the bottom of the modal, it states "Condiciones de utilización: imágenes bajo licencia CC-BY-NC-SA" and shows the Creative Commons license logo.

The background website header includes the URL www.bne.es/ca/Catalogos/BibliotecaDigitalHispanica/Colecciones/, the logo of the Biblioteca Digital Hispánica, and the text "BIBLIOTECA DIGITAL HISPÁNICA BIBLIOTECA NACIONAL DE ESPAÑA". A search bar contains the text "Libres, manuscrits, partitures, fotografies..." and a "CERCAR" button. Navigation tabs include "Inici", "Descobriu col·leccions", and "Sobre la digitalització".

The main content area is titled "COL·LECCIONS DESTACADES" and features a grid of 12 circular thumbnails, each with a caption and a right-pointing arrow:

- Llistes de reproducció de sonsors »
- Art general »
- Atlas i material cartogràfic »
- Cantorals »
- Cartells publicitaris »
- Cartes nàutiques »
- Cervantes »
- Cilindres de cera »
- Dibuixos d'arquitectura »
- Dibuixos dels nens de la Guerra »
- Discos perforats »
- Ephemera »



ありがとう
 Спасибо
 Tak
 d'akujem
 Grazie
 Dankie
 Pakka pér
 matur nuwun
 Danke
 Paldies
 ačiū
 dėkuji
 Dziękuję
 aitäh
 Thank you!
 obrigado
 Tack
 謝謝
 Mauruuru
 Gracias
 hvala
 Dankon
 kiitos
 Na gode
 Gràcies
 Merci
 Grazzi
 ju faleminderit
 eskerrik asko
 terima kasih
 köszönöm