Zoom link https://emory.zoom.us/j/94682801585

Spring 2022 ***Data Science for Beginners*** Soc/Ling/QTM 190

"[an] imagined future in which the long-established way of doing scientific research is replaced by computers that divulge knowledge from data at the press of a button…"

Emory University

Roberto Franzosi

Office Room No.	212 Tarbutton Hall
Email	rfranzo@emory.edu
Office Hours	Tu-Th 1:00-2:30 or by appointment Personal meeting room (Office hours) https://emory.zoom.us/j/8166581703
Lectures	Tu-Th 4:00-5:15 Callaway Center S101

DEADLINES AND IMPORTANT DATES

First day of class	January 11
Last day of class	April 25
Spring break	**March 7-11 no classes**
Software Problems	Raise issue on GitHub
Presentations	Due each Thursday for selected teams, starting on week 3

Table of Contents

COURSE OBJECTIVES

Tools of analysis and visualization of text data

The course deals with new Natural Language Processing (NLP) tools of analysis of text data and visualization (e.g., network graphs, geographic maps). Many of these tools have been developed in conjunction with new technologies of machine learning and Artificial Intelligence aimed at large text corpora available on the web. It is these huge amounts of (mostly textual) data that offer both humanities and social sciences new avenues of research in the form of digital humanities, and where different types of data can be pulled together on a topic and displayed on the internet in very creative ways.

Learning the language of Natural Language Processing (NLP)

From sentence splitter, to tokenizer, lemmatizer, parser with its Part-of-Speech tags (POSTAG), Dependency Relations (DEPREL), Named Entity Recognition (NER), semantic trees, sentence complexity and text readability, noun and verb analysis, n-grams viewer, sentiment analysis, topic modelling, extraction of SVOs (Subject-Verb-Object), and "shape" of stories… you will learn the language of Natural Language Processing (NLP).

The course will show how to use different tools of data visualization, especially **network graphs** dealing with relationships between objects (social actors, concepts, or just words), both static and dynamic (changing with time), and **spatial maps** dealing with objects in space (and time, dynamic maps) through Geographic Information System (GIS) tools.

Big data/small data

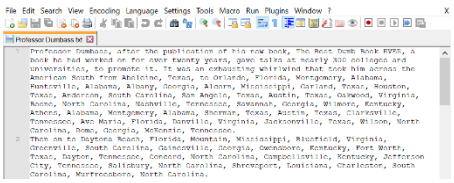
Although the tools used in the course have been developed for big data, the course will mostly deal with small data (e.g., tens of documents) since we do not have the computing power to deal with huge amounts of data.

Visualization and a world of beauty. A game changer?

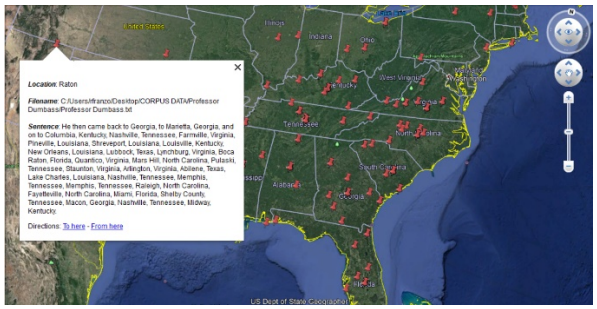
Beyond the technical aspects of data visualization, the course addresses broader questions about **the impact of big data on scholarly practice**. What is the relationship between macro and micro? Does it still make sense to talk about statistical outliers and their role when millions of data points (words) are now used? **Are the new forms of data visualization simply descriptive?** What happened to social sciences' central concern with hypothesis testing?

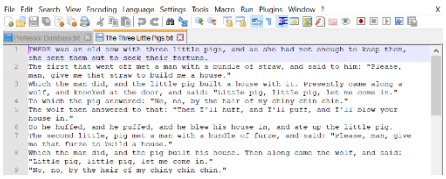
And if color, form, movement, in Kandinsky's view, are the distinctive weapons of art (and beauty), are the new visualization techniques – all based on color, shape, and movement –**are these NLP tools a game changer** in the traditional ways of displaying evidence (i.e., a table of numeric estimate values)? Does this offer a rapprochement between the humanities and science, in approaches, in techniques, perhaps even in modes of writing?

To make a long story short, we basically want to go **automatically,** at the click of a button…



from here (a text file) to here (a map)





from here (a text file) to here (a word cloud)

… ultimately turning words into works of art? IS THIS A GAME CHANGER OF THE NEW DATA SCIENCE?

WHY SHOULD YOU TAKE THIS COURSE: LEARNING OUTCOMES

Welcome to the 21st century!

Have you ever wondered how your smart phone can ask you if you want to call the number or get directions when a friend's text message has a phone number or a city in it? Have you ever wondered how that same smart phone can understand you when you talk to it, whether to ask questions or to dictate to it? And most of the time it even gets it right! This is your 21st century world, a world you are well familiar with. By taking this course, you will get a glimpse at what makes this possible.

Learning outcomes

By the end of term, you will be able to:

1. Understand the concepts of big data, Natural Language Processing (NLP), Artificial

Intelligence, machine learning…
2. Use a variety of NLP tools and what they can do
3. Use a variety of data visualization tools, drawing geographic maps, network graphs, charts …
4. Make public presentations before an audience
5. Write research reports

Ongoing measurement of learning outcomes

Learning outcomes will be assessed every week through weekly presentations.

IS THIS A COURSE FOR YOU?

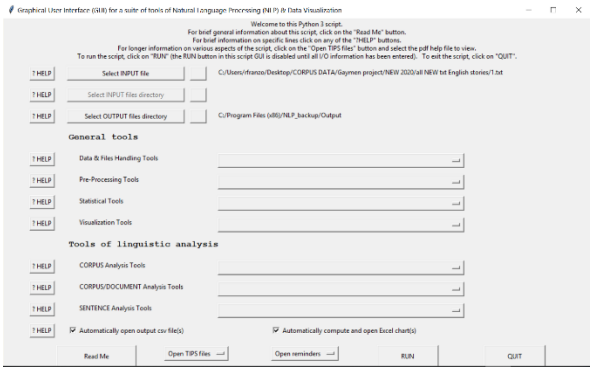
No prerequisites

There are no formal prerequisites for the course, except for a general GOOD familiarity with (and lack of fears of) computers. If you do have a computer science background, of course, you will be able to do more and get more out of the course. But such background is not necessary. In fact, the course was designed with a student in mind with no such background. If you are an Apple user and do not know what the C: drive or the Program files folder is … then, this course may be challenging at the beginning. But one of the best final papers that I have read coming out of this course was written by just such a student!

GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders

After all… All NLP tools in the Suite come with easy-to-use graphical user interfaces (GUI) that make your life easy, with on-line HELP, Read Me messages, reminders and extensive TIPS.

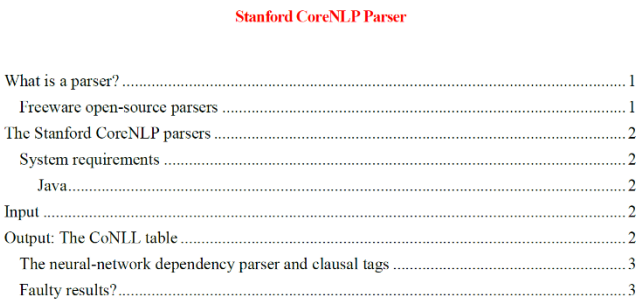
All you need to do is press buttons! **If you know how to do that, you are halfway there…**



The introductory Graphical User Interface (GUI) to the NLP Suite

HELP, Read Me, Videos, TIPS, Reminders buttons are all at your fingertips. **Hard to screw up!**

Example of TIPS file… TIPS files, at least the longer ones, even come with a Table of Contents.



Stanford CoreNLP Parser

Download and install the NLP Suite and other software

From GitHub (https://github.com/NLP-Suite/NLP-Suite/wiki) you need to download and install the NLP Suite appropriate for your machines, Mac or Windows. **You must have a free GitHub account. Please, register on GitHub if you are not already registered. Follow the instructions on the Wiki page of the GitHub NLP Suite.**

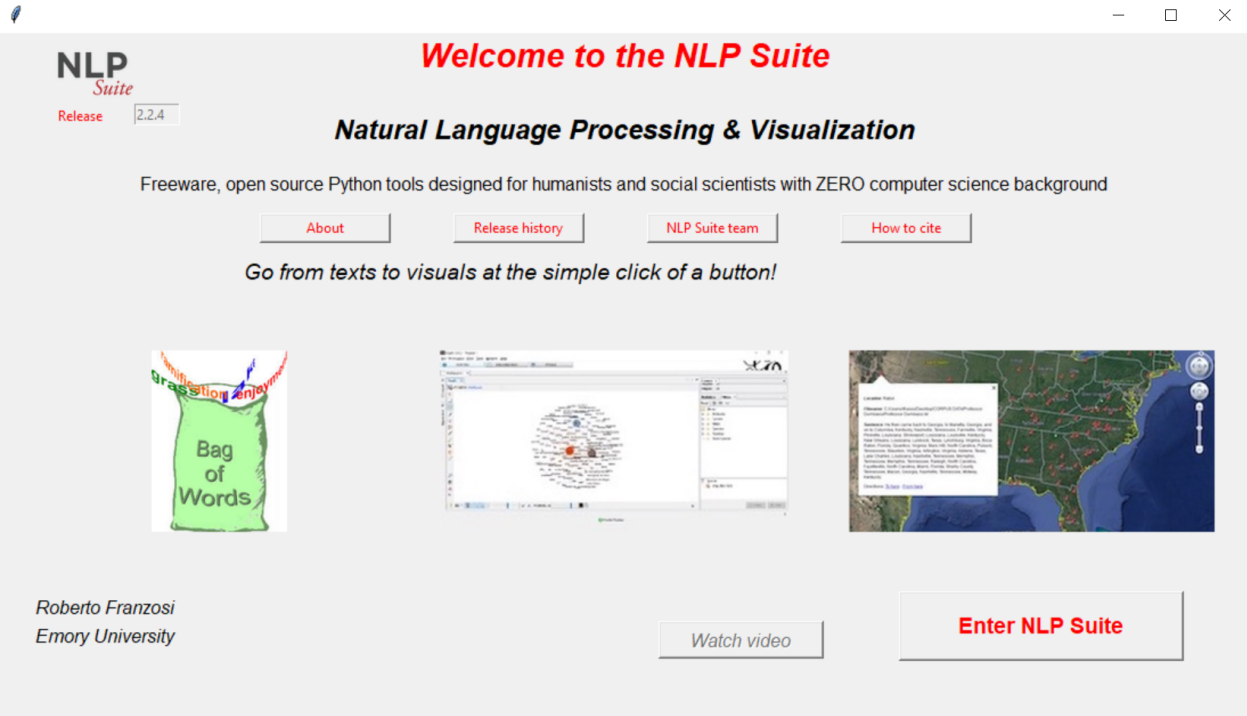
The NLP Suite will automatically install much of the software you need, in particular Python, Anaconda, the NLP Suite, and other Java components.

You will also need to download external software required to run the NLP Suite: JAVA JDK, Stanford CorNLP, SENNA, MALLET, WordNet, Gephi, Google Earth Pro

Please, read carefully all installation instruction in the wiki of the NLP Suite GitHub repository.

NLP Suite welcome GUI

Once installed, you can run the NLP Suite that will open the following welcome GUI



You need a work partner

Undergraduates in the class will work with a partner, in teams of **2 students per team**. **Each team will last through the semester, starting on week 3.**

You are welcome to choose your own partner, otherwise, after week three we will randomly assign students to teams. And to keep honest people honest, on each presentation you need to state the % contribution of each partner.

You need a corpus

We have several text corpora that you can analyze.

Gay men project

376 personal narratives from gay men from 37 different countries

1. **The Harry Potter books**
   J. K. Rowling's collection of 7 Harry Potter books
2. **US presidential speeches (https://www.presidency.ucsb.edu/)**
   a. Inaugural addresses
      A collection of 62 inaugural addresses by US presidents (1789-2021)
   b. State-of-the-union addresses
      A collection of 234 state-of-the-union addresses by US presidents (1790-2021)
3. ***New York Times* best-selling book reviews**
   Some 1300 NYT best-selling book reviews
4. **Folktales**

A collection of hundreds of cross-national English, German, Chinese, Arabic, and Indian folktales

Homework (due Sunday January 16, at midnight): Installing NLP software for distant reading

Provide screenshots of successful installation of software on your computer.

1. What do those batch files in setup_Mac or setup_Windows do?
2. How can you make sure that you are always working with the most recent release of the NLP Suite on GitHub according to the GitHub wiki pages?
3. When you open the NLP Suite to run a specific script, the script warns you that you are missing a Python package and that you need to pip install it. You do so. Installation was successful. You run the NLP Suite again. You get the same error. Why? Where would you have gone wrong according to the GitHub wiki pages?

Weekly presentations

The **weekly presentations**, by and large, consist of analyzing text corpora. Each week, students are expected to analyze their corpora using different Natural Language Processing (NLP) tools and to write up the results of their analyses, submitting their work in the form of a Word document. This document will include figures with the results of the NLP analyses (typically, screenshots of computer output) and the students' interpretations and explanations of these figures. What do the results mean? What do they tell you about the substance of the texts? What are the limits of the tools used? On average, some pages of writing are expected every week. But the amount of writing may increase week after week as students return to the same texts using different approaches and tools, ultimately incorporating all of their analyses into one document as they approach submission of the final paper. **Students are also expected to ground their analyses in the body of scholarly literature and TIPS assigned as required readings.**

Presentation rubrics

Each assignment is graded (0-100) and comments are provided. **Weekly rubrics for the presentations are also provided**, detailing the scale for different points. Every week, you will know exactly what you missed! **Rubrics are posted under Files on CANVAS.** Rubrics only serve as a guideline. Gross errors of interpretation of data results or of basic understanding of the tools will be marked down regardless of rubric.

GRADING

Grading will be based on the following items:

Participation (25%). You are expected to attend classes regularly (attendance is enforced through a sign-up sheet) and contribute to discussion.

Presentations (75%) – Starting on week 3 student teams will make in-class presentation of their work. 10-15 minutes max in Power Point with the use of graphical displays.

Presentations will cover an overview of the corpus (what is the corpus about? number of documents, of sentences per document, linguistic domain as shown by the distribution of words) and the most significant results using the tools learned by the time of the presentation (from n-grams, to topic modeling, CoreNLP annotators – gender, normalized dates, quote – knowledge-base and dictionary annotators, SVO extractors, sentiment, style, and more… What are the pros and cons, strengths and limits of the NLP tools used? **As the semester progresses and students learn more NLP tools, repeated teams' presentations are expected to provide both broader and more in-depth analyses of the corpora.**

Bonus points. Students who would like to earn bonus points can write TIPS files that we do not have (or improve files we do have). **Bonus points will be used to help students who are borderline between final grades.**

Students who are not satisfied with a grade received are welcome to ask for re-grading for well-motivated reasons. The result of re-grading may be a higher grade, the same grade, or a lower grade.

HONOR CODE

The Emory University honor code applies fully to this course. When you sign an exam or submit your assignments, you are pledging to the honor code. For reference, please consult: http://catalog.college.emory.edu/academic/policies-regulations/honor-code.html

WEEKLY TOPICS & READINGS

Required & suggested readings

The syllabus lists a number of readings, books and articles. **You are responsible for the required readings only.** Suggested readings are there as bibliographical references in case you want to pursue some topics further. **For the purpose of your grade, you are not expected to read suggested readings** (unless, of course, you are a glutton for punishment! Although … it is also true that the more you read, the more you know… and the better you would do in your presentations and written work).

Where will you find the readings?

All readings, including most of the suggested readings, are uploaded to CANVAS as a downloadable zip file. The readings are **not on Ereserve!!!**

Introduction (Week 1, January 11-13)

Big data and "distant reading"
Digital humanities: What is it?
NLP: What is it?
Becoming familiar with the suite of Java and Python NLP tools

Required readings:

TIPS_NLP_Things to do with words NLP approach.pdf

Brownlee, Jason. 2020. "What Is Natural Language Processing?" retrieved 12/20/2020 https://machinelearningmastery.com/natural-language-processing/

Caliskan, Aylin, Joanna J. Bryson, Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science*, Vol. 356, pp. 183–186.

"Meet GPT-3. It Has Learned to Code (and Blog and Argue)" *The New York Times,* 11/25/2020.

Franzosi, Roberto. 2020. "What's in a Text? Bridging the Gap between Quality and Quantity in the Digital Era." *Quality & Quantity*, DOI 10.1007/s11135-020-01067-6.

Kirschenbaum, Matthew G. 2012. "What is Digital Humanities and What's it Doing in English Departments?" In: pp. 3-11, Matthew K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Video. 13 minutes. Talk by Nello Cristianini. "The Story of Don Antonio." https://youtube.com/seeapattern

Suggested readings:

Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review*, Vol. 54, Vol. 1, pp. 54-68.

Underwood, Ted. 2016. "Distant Reading and Recent Intellectual History." In: pp. 530-533, Matthew K. Gold and Lauren F. Klein (eds.), *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press.

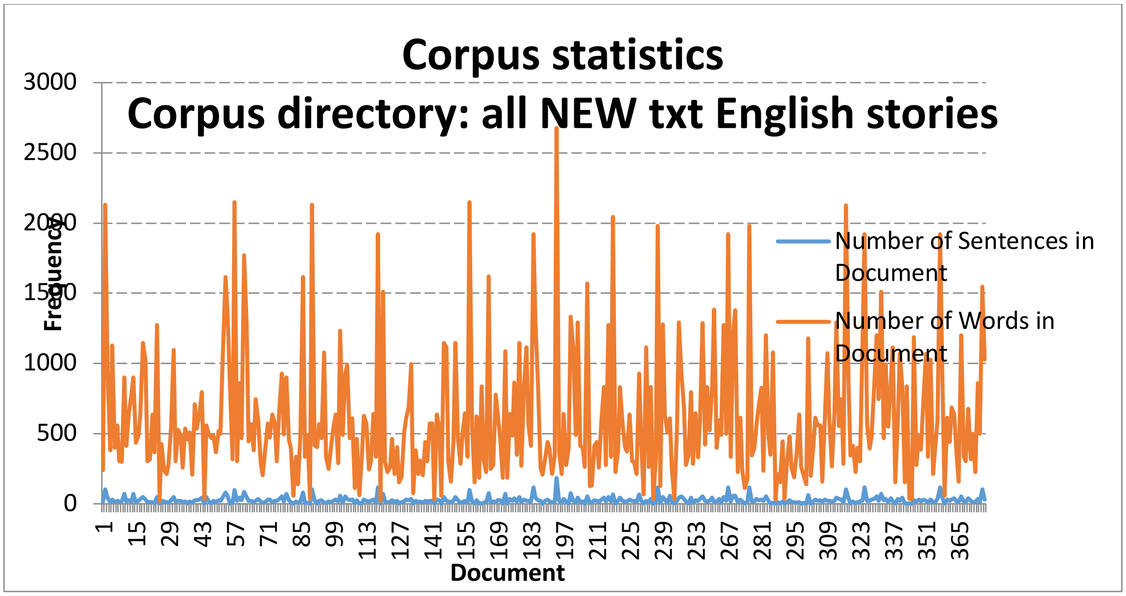
Digital humanities websites: Trans-Atlantic Slave Trade (http://www.slavevoyages.org) by David Eltis, Georgia Civil Rights Cold Cases (https://scholarblogs.emory.edu/emorycoldcases) by Hank Klibanoff

The Digital Scholarship Lab at the University of Richmond, http://dsl.richmond.edu/
The Yale photographic site http://photogrammar.yale.edu/ for the visualization of some 170,000 photographs from 1935 to 1945 created by the United States Farm Security Administration and Office of War Information (FSA-OWI).

Atlas of Early Printing at the University of Iowa, http://atlas.lib.uiowa.edu

Part I (Week 2, January 18-20): Corpus Statistics and Words Visualization

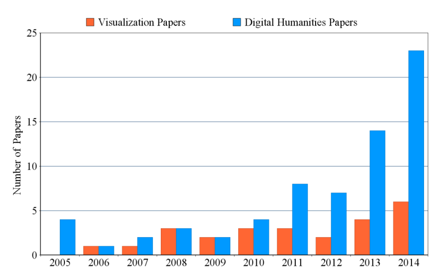
File types doc, docx, rtf, txt, pdf) and what to do about it
Corpus Statistics
Get basic statistics about your corpus: number of documents, number of sentences, number of words; Ngrams



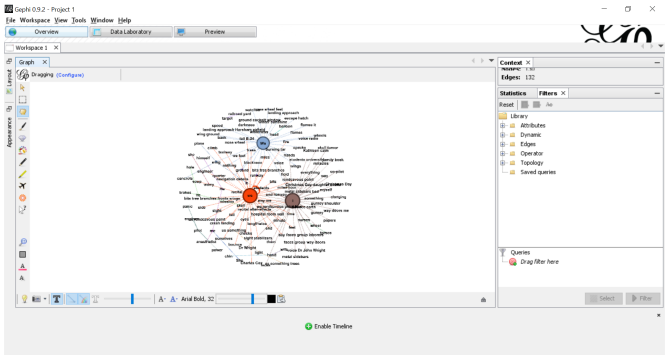
Visualization in Digital humanities
Word clouds

Software: Bookworm, Wordle, TagCrowd, Tagul (now renamed WordArt) and Tagxedo (Tagul and Tagxedo allow to draw word clouds in specific shapes)

Excel charts (with hover-over effects)

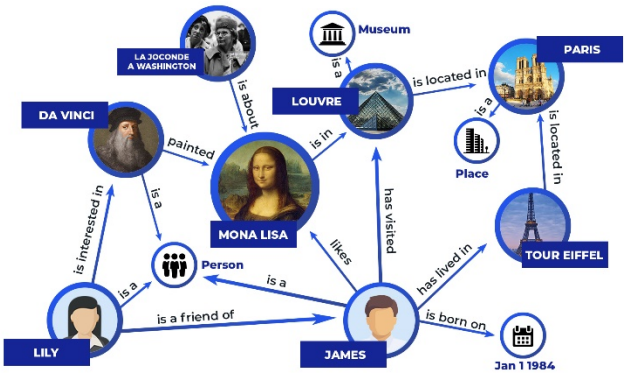


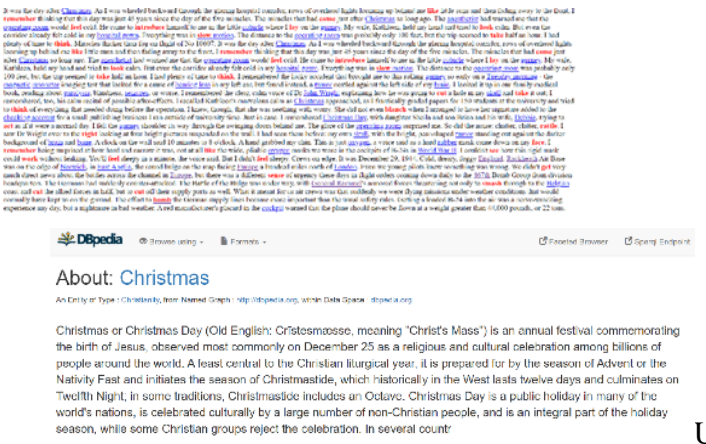
Network graphs: Mapping relations



Software: Gephi

Knowledge graphs (KG) and HTML annotated files





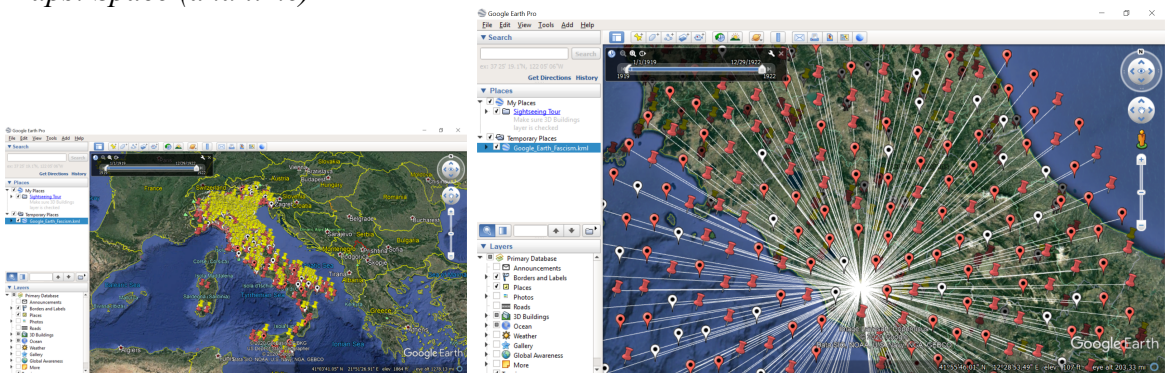
DBpedia

About: Christmas

Christmas or Christmas Day (Old English: Cristesmæsse, meaning "Christ's Mass") is an annual festival commemorating the birth of Jesus, observed most commonly on December 25 as a religious and cultural celebration among billions of people around the world. A least central to the Christian liturgical year, it is prepared for by the season of Advent or the Nativity Fast and initiates the season of Christmastide, which historically in the West lasts twelve days and culminates on Twelfth Night; in some traditions, Christmastide includes an Octave. Christmas Day is a public holiday in many of the world's nations, is celebrated culturally by a large number of non-Christian people, and is an integral part of the holiday season, while some Christian groups reject the celebration. In several countr

Using DBpedia to annotate Murphy's text and clicking on an annotated word (e.g., Christmas) in the html output to access DBpedia.

Maps: Space (and time)



Software: Google Earth Pro, Google Maps

Required readings:

TIPS_NLP_Text encoding.pdf
TIPS_NLP_Text encoding (utf-8).pdf
TIPS_NLP_File checker & converter & cleaner.pdf

TIPS_NLP behind the whats' in your corpus and wordclouds GUIs
The 8 Best Free Word cloud Creation Tools for Teachers: http://elearningindustry.com/the-8-best-free-word-cloud-creation-tools-for-teachers
Nine free on-line word clouds generators: http://www.smashingapps.com/2011/12/15/nine-excellent-yet-free-online-word-cloud-generators.html

Video. 14 minutes. Ted Talk by Erez Lieberman Aiden and Jean-Baptiste Michel, 2011, "A picture is worth 500 billion words". https://www.youtube.com/watch?v=WtJ50v7qByE&t=19s

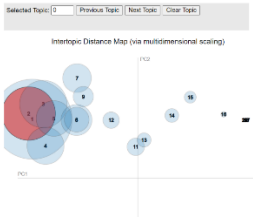
Goldstone, Andrew and Ted Underwood. 2014. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History*, Vol. 45, No. 3, pp. 359-384.
Heimerl, Florian, Steffen Lohmann, Simon Lange, and Thomas Ertl. 2014. "Word cloud explorer: Text analytics based on word clouds." *47th Hawaii International Conference on System Sciences*. IEEE, 2014.

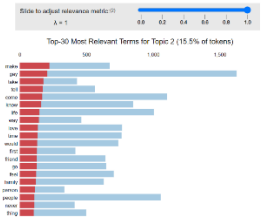
Suggested readings:

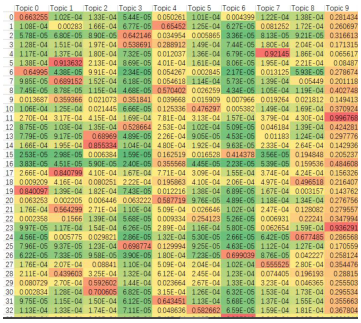
Corman, Steven R., Timothy Kuhn, Robert D. Mcphee, and Kevin J. Dooley. 2002. "Studying Complex Discursive Systems." *Human Communication Research*, 28(2):157–206.
Wilkinson, Leland and Michael Friendly. 2009. "The History of the Cluster Heat Map." *The American Statistician*, Vol. 63, No. 2, pp. 179-184.

Part II (Week 3, January 25-27): Topic Modeling & Word2Vec

What are the topics in your corpus?
Topic modeling via Gensim and Mallet
Word2Vec







Software: Mallet & Gensim

Franzosi, Roberto. NLP TIPS files.

Graham, Shawn, Scott Weingart and Ian Milligan. 2012. *Getting Started with Topic Modeling and MALLET*. The Programming Historian. Document available on the web at http://programminghistorian.org/lessons/topic-modeling-and-mallet
For an interesting paper based on Gensim and with various practical recommendations and references, see:

Micah Saxton's Capstone. *Topic Modeling Best Practices*. https://msaxton.github.io/topic-model-best-practices/

Suggested readings:

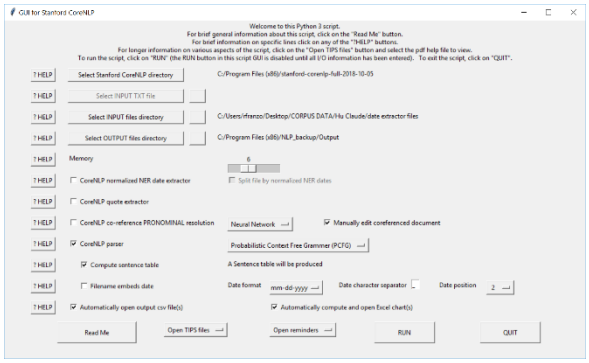
There are some great readings in this 2013 special issue of *Poetics*. Take a quick look at these articles and dive deeper in the ones that go to the heart of your interests.

DiMaggio, Paul, Manish Nag, and David Ble. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics*, Vol. 41, No. 6, pp. 570-606.
Marshall, Emily A. 2013. "Defining Population Problems: Using Topic Models for Cross-national Comparison of Disciplinary Development," *Poetics*, Vol. 41, No. 6, pp. 701-724.
McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu.
McFarland, Daniel A. and Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, Daniel Jurafsky. 2013. "Differentiating Language Usage through Topic Models," *Poetics*, Vol. 41, No. 6, pp. 607-625.
Miller, Ian Matthew. 2013. "Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach," *Poetics*, Vol. 41, No. 6, pp. 626-649.
Tangherlini, Timothy R. and Peter Leonard. 2013. "Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research," *Poetics*, Vol. 41, No. 6, pp. 725-749.
Flaounas, Ilias, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini, 2013 "Research Methods in the Age of Digital Journalism: Massive-scale Automated Analysis of News: Content Topics, Style and Gender," *Digital Journalism*, Vol. 1, No. 1, pp. 102–116.
Martin, Fiona and Mark Johnson. 2015. "More Efficient Topic Modelling Through a Noun Only Approach." *Proceedings of the Australasian Language Technology Association Workshop*, pp. 111−115.

Video on the differences between Artificial Intelligence, Machine Learning, and Deep Learning
https://www.youtube.com/watch?v=WSbgixdC9g8

Part III (Week 4, February 1-3): NLP (Natural Language Processing): Basic language

Sentence splitter, tokenizer, lemmatizer, parser
The Stanford CoreNLP parsers
Meet the CoNLL table



Software: Stanford CoreNLP

Required readings:

Top 20 free software for Text Analysis, Text Mining, Text Analytics
http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

Franzosi, Roberto. NLP TIPS files.

Video. 14 minutes. Talk by Nello Cristianini on Big Data ("Patterns in Media Content)
https://www.youtube.com/watch?v=mmWRNRPb0W0

Suggested readings:

Take a quick look at some of these readings. Familiarize yourself with what the ready availability of digital newspaper archives would allow you to do/and how.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences (PNAS), PNAS,* Published online January 9, 2017 E457–E465.

Sudhahar, Saatviga, Giuseppe A. Veltri, and Nello Cristianini. 2015. "Automated Analysis of the US Presidential Elections Using Big Data and Network Analysis." *Big Data & Society*, Vol. 2, No. 1, pp. 1–28. DOI: 10.1177/2053951715572916.

Mohr, John, Robin Wagner-Pacifici, Ronald L. Breiger, Petko Bogdanov. 2013. "Graphing the Grammar of Motives in National Security Strategies - Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics," *Poetics*, Vol. 41, No. 6, pp. 670-700.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data," *Theory and Society*, Vol. 43, No. 3, pp. 465-482.

Luhn, H.P. 1959. "Keyword in Context Index for Technical Literature (KWIC Index)." Yorktown Heights, NY: IBM, Report RC 127. Also in: 1960. *American Documentation*, Vol. 11, pp. 288–295.

Seguin, Charles. 2015 web download. "Scraping Historical Newspaper Archives: The Transformation of Public Lynching Inquiry." Sociological Theory 10:164–93.

Light, Ryan. 2014. "From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses." *Social Currents*, Vol. 1, No. 2, pp. 111–129.

Light, Ryan and Jeanine Cunningham. 2016. "Oracles of Peace: Topic Modeling, Cultural Opportunity, and the Nobel Peace Prize, 1902–2012." *Mobilization: An International Quarterly*, Vol. 21, No. 1, pp. 43–64.

He, Qin. 1999. "Knowledge Discovery through Co-word Analysis." *Library Trends* 48:133–59.

Discourse in the US." http://badhessian.org/2014/01/scraping-historical-newspaper-archives-the-transformation-of-public-lynching-discourse-in-the-us/Snowsill, Tristan, Ilias Flaounas, Tijl De Bie, and Nello Cristianini. 2010. "Detecting Events in a Million New York Times Articles," *Lecture Notes in Computer Science,* pp. 615-618.

Zervanou, Kalliopi, Marten Düring, Iris Hendrickx, and Antal van den Bosch. 2014. "Documenting Social Unrest: Detecting Strikes in Historical Daily Newspapers," *Lecture Notes in Computer Science,* pp. 120-133.

Part IV (Week 5, February 8-10): Named Entity Recognition (NER) and CoreNLP annotators

*A closer look at the CoNLL table**:** Meet the NER, POSTAG, DEPREL tags*
Stanford CoreNLP annotators
Is there dialogue?
Are there people and organizations and differences in gender distribution?

Use CoreNLP NER annotator and gender annotator, and the names databases

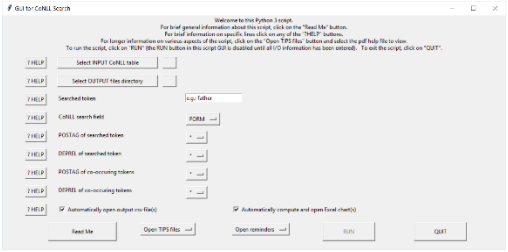
Are there geographical locations?

Use CoreNLP NER annotator to extract geocodable locations (COUNTRY, STATE OR PROVINCE, CITY) and informal locations (LOCATION)

Use WordNet to get lists of both proper geographic locations and improper locations (kitchen)

Are there times?

Use CoreNLP NER normalized time annotator to extract standardized temporal expressions



Software: Stanford CoreNLP

Using WordNet: Does nature appear?

Use WordNet (noun synsets plant, animal; verb synset weather) to get listings of animals, plants, and weather)

Using WordNet: Do nouns and verbs cluster in specific classes?

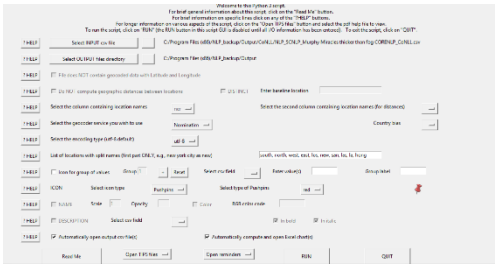
Use WordNet to aggregate verbs and nouns in your corpus and compute frequency distributions of classes.

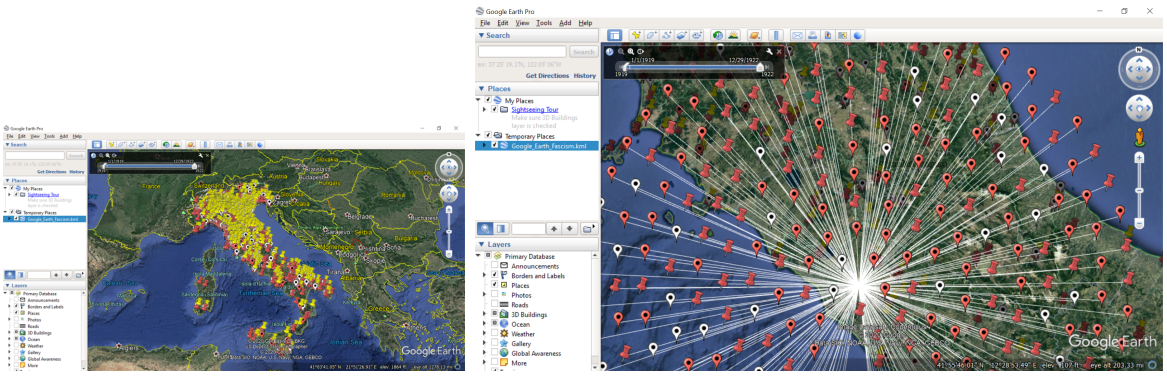
Required readings:

Franzosi, Roberto. NLP TIPS files.

Part V (Week 6, February 15-17): From text to maps

Using CoNLL NER information to map locations
Geocoding
Visualizing time and space





Software: Carto, Google Earth Pro, QGIS, Tableau, TimeMapper, GeoNames, OpenStreetMap

Required readings:

Franzosi, Roberto. Geocoding TIPS files.

Graham, Mark and Taylor Shelton. 2013. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography*, Vol. 3, No. 3, pp. 255–261.
Lewis, Peirce. 1985. "Beyond Description." Annals of the Association of American Geographers, Vol. 75, No. 4, pp. 465-478.Yuan, May. 2010. "Mapping Text". In: pp. 109-123, David J. Bodenhamer, John Corrigan, and Trevor M. Harris (eds.), The Spatial Humanities: GIS and the Future of Humanities Scholarship. Bloomington, IN: Indiana University Press.

Suggested readings:

Basso, Keith H. 1988. "'Speaking with Names': Language and Landscape among the Western Apache." *Cultural Anthropology*, Vol. 3, No.2, pp. 99-130.
Corrigan, John. 2010. "Qualitative GIS and Emergent Semantics". In: pp. 76-88, David J. Bodenhamer, John Corrigan, and Trevor M. Harris (eds.), The Spatial Humanities: GIS and the Future of Humanities Scholarship. Bloomington, IN: Indiana University Press.
Gregory, Ian, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. "Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research." *International Journal of Humanities and Arts Computing*, Vol. 9, No. 1, pp. 1–14.
Jessop, Martyn. 2008. "The Inhibition of Geographical Information in Digital Humanities Scholarship." *Literary and Linguistic Computing*, Vol. 23, No. 1, pp. 39-50.
Kitchin, Rob. 2013. "Big Data and Human Geography: Opportunities, Challenges and Risks." *Dialogues in Human Geography*, Vol. 3, No. 3, p. 262–267.
Massey, Doreen. 2005. *For Space.* Thousand Oaks, CA: Sage.
Ó Murchú T. and S. Lawless. 2014. "The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data." In *Proceedings of the Digital Humanities*. (2014). 7, 8, 12.
Rosenberg, Daniel and Anthony Grafton. 2010. *Cartographies of Time*. New York, Princeton Architectural Press.
Yuan, May. 2014. "Temporal GIS for Historical Research." In: pp. 45-55, A. Crespo Solana (ed.), Spatio-Temporal Narratives: Historical GIS and the Study of Global Trading Networks. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Check out some cool mapping sites

http://www.radicalcartography.net/
http://selfiecity.net/
http://www.floatingsheep.org/

http://dsl.richmond.edu/
http://photogrammar.yale.edu/
http://atlas.lib.uiowa.edu

Part VI (Week 7, February 22-24): Narrative and the 5 Ws

The 5 Ws of Narrative: Who does What, When, Where, and Why

SVO Extraction & Visualization
Stanford CoreNLP enhanced dependencies parser
SENNA
Stanford CoreNLP OpenIE

Computer scientists are coming closer to finding automated solutions to extracting the "who, what, when, where, why, and how" of narrative. It will not be long before they will put social scientists out of their miseries of manual coding!

Required readings:

Franzosi, Roberto. NLP TIPS files.

Franzosi, Roberto. 2012. "On Quantitative Narrative Analysis." In: pp. 75-98, James A. Holstein and Jaber F. Gubrium (eds.), *Varieties of Narrative Analysis*. Thousand Oaks, CA: Sage.

Franzosi, Roberto, Wenqin Dong, Ziyang Hu, and Gabriel Wang. 2020. "*Automatic Information Extraction of the Narrative Elements Who, What, When, and Where*." Paper under journal review.

Suggested readings:

Chambers, Nathanael and Dan Jurafsky. 2010. "A Database of Narrative Schemas." In *Proceedings of LREC-2010*, Palo Alto, CA, USA, 2010.

Del Corro, Luciano and Rainer Gemulla. 2013. "ClausIE: Clause-Based Open Information Extraction." Proceeding WWW '13 *Proceedings of the 22nd international conference on World Wide Web*, pp. 355-366, Rio de Janeiro, Brazil – May 13-17, 2013.

Lansdall-Welfare, Thomas and Nello Cristianini. 2020. "History Playground: A Tool for Discovering Temporal Trends in Massive Textual Corpora". *Digital Scholarship in the Humanities*, Vol. 35, No. 2, pp. 327-341. http://playground.enm.bris.ac.uk

John, Markus, Steffen Lohmann, Steffen Koch, Michael Wörner, and Thomas Ertl. 2016. "Visual Analysis of Character and Plot Information Extracted from Narrative Text." In: pp. 220-241, Braz, José February, Nadia Magnenat-Thalmann, Paul Richard, Lars Linsen, Alexandru Telea, Sebastiano Battiato, Francisco Imai (Eds.). *Computer Vision, Imaging and Computer Graphics Theory and Applications 11th International Joint Conference, VISIGRAPP 2016* Rome, Italy, 27–29, 2016. Cham, Switzerland: Springer.

Finlayason, Mark Alan. 2012. *Learning Narrative Structure from Annotated Folktales.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

John, Markus, Martin Baumann, David Schuetz, Steffen Koch, and Thomas Ertl. 2019. "A Visual Approach for the Comparative Analysis of Character Networks in Narrative Texts," *2019 IEEE Pacific Visualization Symposium (PacificVis), Bangkok, Thailand, 2019*, pp. 247-256.

Ó Murchú, Tomás and Séamus Lawless. 2014. "The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data." In *Proceedings of the Digital Humanities*. 7, 8, 12.

(found under Murchú or zip will not zip)

Lendvail, Piroska, Thierry Declerck, Sándor Darányi, Pablo Gervás, Raquel Hervás, Scott Malec, and Federico Peinado. 2010. "Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case," *Proceedings of the Seventh conference on International Language Resources and Evaluation, European Language Resources Association (ELRA)*.

Palmer, Martha and Daniel Gildea. 2004. "The Proposition Bank: An Annotated Corpus of Semantic Roles." *Computational Linguistics*, Vol. 20, No. 10, pp. 1-33.

Palmer, Martha. 2008. Propbank, A corpus annotated with semantic roles." http://verbs.colorado.edu/~mpalmer/dossier/HindiIntro.pdf

Scott Malec, Sándor Darányi, Trevor Cohen, and Dominic Widdows. [no date]. "Landing Propp in Interaction Space: First Steps toward Scalable Open Domain Narrative Analysis with Predication-based Semantic Indexing."

Sudhahar, Saatviga, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. 2015. "Network Analysis of Narrative Content in Large Corpora," *Natural Language Engineering*, Vol. 21, No. 1, pp. 81-112.

Sudhahar, Saatviga and Nello Cristianini. 2013. "Automated Analysis of Narrative Content for Digital Humanities," *International Journal of Advanced Computer Science*, Vol. 3, No. 9, Pp. 440-447.

Sudhahar, Saatviga, Thomas Lansdall-Welfare, Ilias Flaounas, and Nello Cristianini. 2012. "Quantitative Narrative Analysis of US Elections in International News Media." The Internet, Policy & Politics Conferences, Oxford Internet Institute, University of Oxford. http://ipp.oii.ox.ac.uk/2012/programme-2012/track-a-politics/panel-5a-topics-memes-and-sentiment/saatviga-sudhahar-thomas-lansdall

Hanna, Alex. 2017. "MPEDS: Automating the Generation of Protest Event Data." SocArXiv Preprints. https://osf.io/preprints/socarxiv/xuqmv/ DOI 10.31235/osf.io/xuqmv.

UzZaman, Naushad, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. "SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations." *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, June 14-15, 2013.

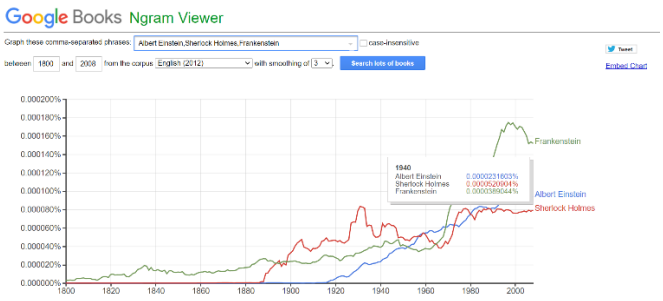
Zhang, Han and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology*, Vol. 49, No. 1, pp. 1–57.

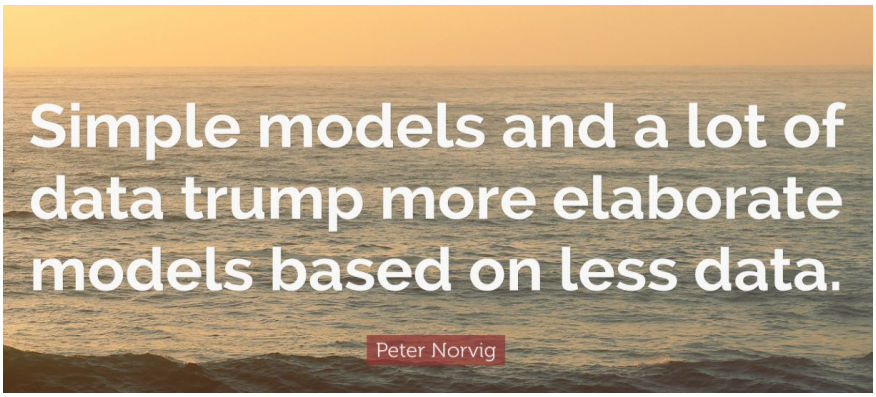
Part VII (Weeks 8-9, March 1-3, March 8-10): N-grams, co-occurrences, culturomics

Week 8: March 8-10

A closer look at N-grams
Google N-grams Viewer and Culturomics
N-grams searches in the NLP Suite
Word co-occurrences searches
Single words/collocations searches

Software: Stanford CoreNLP, Google Ngram Viewer





Required readings:

Franzosi, Roberto. NLP TIPS files.

Become familiar with the basic language of culturomics!

Video. 14 minutes. Ted Talk by Erez Lieberman Aiden and Jean-Baptiste Michel, 2011, "A picture is worth 500 billion words". https://www.youtube.com/watch?v=WtJ50v7qByE&t=19s

Same Video. 14 minutes. Michel, Jean-Baptiste and Erez Lieberman Aiden. 2011. "What we learned from 5 million books". https://www.ted.com/talks/what_we_learned_from_5_million_books?language=en

Anderson, Chris. 2008. "The end of theory: The data deluge makes the scientific method obsolete." *Wired Magazine*, Vol. 16, No. 7,

Available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Mazzocchi, Fulvio. 2015. "Could Big Data be the End of Theory in Science?" *EMBO Reports*, Vol. 16, No. 10, pp. 1250-1255. doi:10.15252/embr.201541001.

Suggested readings:

De Marneffe, Marie-Catherine, and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. Technical report, Stanford University, 2008.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 14 January 2011, Vol. 331, pp. 176-182.

Letcher, David W. 2011. "Cultoromics: A New Way to See Temporal Changes in the Prevalence of Words and Phrases." *American Institute of Higher Education 6th International Conference Proceedings*. Vol. 4, No.1, pp. 228-236.

Leetaru, Kalev H. 2011. "Culturomics 2.0: Forecasting Large–scale Human Behavior Using Global News Media Tone in Time and Space." *First Monday,* Vol. 16, No. 9 (on-line journal).

Nunberg, Geoffrey. 2009. "Google's Book Search: A Disaster for Scholars." *The Chronicle of Higher Education,* August 31, 2009.

Schwartz, Tim. 2011. "Culturomics Periodicals Gauge Culture's Pulse." *Science,* Vol. 332, 1 April 2011, p. 35-36.

Jurafsky, Daniel and James H. Martin. 2020. "N-gram Language Models." *Speech and Language Processing*. Available online at https://web.stanford.edu/~jurafsky/slp3/

Week 9: March 1-3

SPRING BREAK March 8-10 no classes

Part VIII (Week 10, March 15-17): Knowledge-base systems (DBpedia and YAGO)

DBpedia
YAGO
Dictionary-based annotation
html files

Required readings:

Franzosi, Roberto. NLP TIPS files.

Huet, Thomas, Joanna Biega, and Fabian M. Suchanek. 2013. "Mining History with Le Monde." *ACM 978-1-4503-2411-3/13/10* http://dx.doi.org/10.1145/2509558.2509567

Ringler, Daniel and Heiko Paulheim. 2017. "One Knowledge Graph to Rule Them All? Analyzing the Differences Between DBpedia, YAGO, Wikidata & co." In: Kern-Isberner G., Fürnkranz J., Thimm M. (eds) *KI 2017: Advances in Artificial Intelligence. KI 2017. Lecture Notes in Computer Science*, vol 10505. Springer, Cham. https://doi.org/10.1007/978-3-319-67190-1_33.

Suggested readings:

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendesf, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. 2012. "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia." *Semantic Web*, Vol. 1, pp. 1–5.

Suchanek, Fabian M. Gjergji Kasneci, and Gerhard Weikum. 2008. "Yago: A Large Ontology from Wikipedia and WordNet." *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*. Vol. 6, pp. 203-217.

Part IX (Weeks 11-12, March 22-24, March 29-31): The world of emotions

Week 11: March 22-24

The words of emotions

You can use WordNet to get lists of all nouns (*feeling* WordNet noun class) and all verbs (*emotion* WordNet verb class) of emotions in the English language.

You can use the YAGO annotator (*Emotion* YAGO class) to get lists of words of emotion found in your specific corpus.

The rhetoric of emotions: punctuation and repetition

The use of question marks and exclamation marks which contribute to the rhetorical figures of speech of pathos. And so does repetition, as part of a figure of amplification.

Sentiment Analysis: Capturing the feelings conveyed in the writing
WordNet
YAGO
ANEW
Hedonometer
SentiWordNet
Stanford CoreNLP sentiment analysis annotator
VADER

Video. Talk by Min Song on Sentiment Analysis. https://www.coursera.org/learn/text-mining-analytics/lecture/5RwtX/5-6-how-to-do-sentiment-analysis-with-sentiwordnet

Franzosi, Roberto. NLP TIPS files.

Suggested readings:

You can download SentiWordNet at http://sentiwordnet.isti.cnr.it/

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche. Pisa, IT.

Bradley, Margaret M. and Peter J. Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. NIMH Center for the Study of Emotion and Attention. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Dodds, Peter Sheridan and Christopher M. Danforth. 2010. "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents." *Journal of Happiness Studies*, Vol. 11, pp. 441–456.

Esuli, Andrea and Fabrizio Sebastiani. 2006. "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining." In: pp. 417–422. *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, Genova, IT.

Nguyen, Thin, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. 2010. "Classification and Pattern Discovery of Mood in Weblogs." In: pp. 283–290, M. J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, Berlin: Springer-Verlag.

Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank." *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas." 2013. *Behavior Research Methods*. Advance Online Publication. DOI: 10.3758/s13428-012-0314-x. [PubMed]

Hills, Thomas T. and James S. Adelman. 2015. "Recent Evolution of Learnability in American English from 1800 to 2000." *Cognition*, Vol. 143, pp. 87–92.

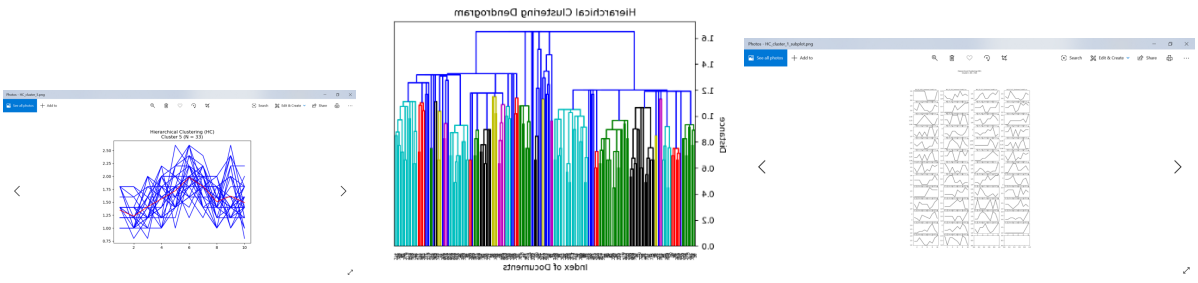
Hills, Thomas, Eugenio Proto, and Daniel Sgroi. 2015. "Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books." *IZA (Forschungsinstitut zur Zukunft der Arbeit/Institute for the Study of Labor)*, Discussion Paper No. 9195, pp. 1-25.

Reagan, Andrew J., Christopher M. Danforth, Brian Tivnan, Jake Ryland Williams, Peter Sheridan Dodds. 2016. "Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs." Download from https://arxiv.org/abs/1512.00531.

Ribeiro, Filipe N., Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. "Sentibench-A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods." *EPJ Data Science*, Vol. 5, No. 1, pp. 1-29.

Week 12: March 29-31

The "shape" of stories
Data reduction algorithms: Hierarchical Clustering (HC), Singular Value Decomposition (SVD), Non-Negative Matrix Factorization (NMF)



Required readings:

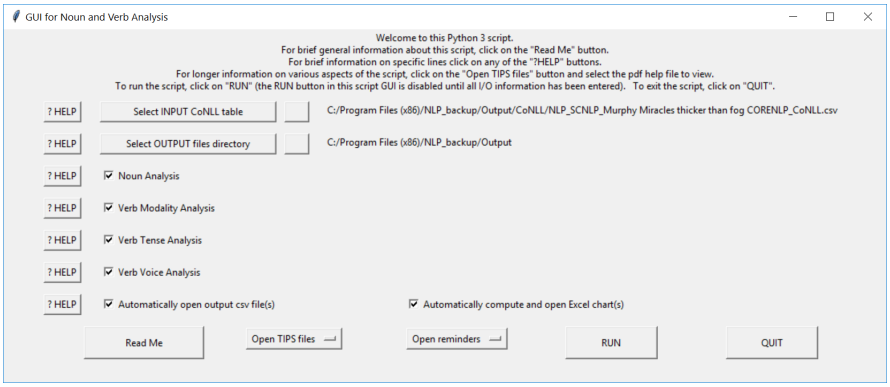
Franzosi, Roberto. NLP TIPS files.

Vonnegut, Kurt. 2005. "Here is a Lesson in Creative Writing." In: pp. 23-28, Kurt Vonnegut, *A Man Without a Country. Edited by Daniel Simon.* New York: Seven Stories Press.
Video. Vonnegut, Kurt. https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Suggested readings:

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". *EPJ Data Science*, Vol. 5, No. 31, pp. 1-12.
Burrows, J.F. 1987. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing*, Vol. 2, No. 2, pp. 61-70.

Part X (Week 13 April 5-7): Dissecting your corpus via the CoNLL table



Noun density and noun types
Verb modality: Ability, possibility, permission, and obligation
Verb tense: past, future, gerundive
Verb voice: Active and passive verb forms
Function words ("junk" words or "stop" words): pronouns, prepositions, articles, conjunctions, and auxiliary verbs
Pronouns and Coreference resolution

Software: Stanford CoreNLP, WordNet

Required readings:

Franzosi, Roberto. NLP TIPS files.

Franzosi, Roberto, Gianluca De Fazio, and Stefania Vicari. 2012. "Ways of Measuring Agency and Action: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875-1930)." In: pp. 1-41, Tim Liao (ed.), *Sociological Methodology*, Vol. 42.
Moretti, Franco and Dominique Pestre. 2015. "BANKSPEAK: The Language of World Bank Reports." *New Left Review*, Vol. 92, pp. 75-99. Moretti
Chung, Cindy and James Pennebaker. 2007. "The Psychological Functions of Function Words." In: pp. 343-359, Klaus Fiedler (Ed.), Social Communication, New York: Psychology Press.

Suggested readings:

Bonyadi, Alireza. 2011. "Linguistic Manifestations of Modality in Newspaper." *International Journal of Linguistics*, Vol. 3, No. 1, E30.
Newman, Matthew L, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. "Lying Words: Predicting Deception from Linguistic Styles." Personality and Social Psychology Bulletin, Vol. 29 No. 5, pp. 665-675.
Flood, Barbara J. 1999. "Historical Note: The Start of a Stop List at Biological Abstracts." *Journal of the American Society for Information Science*, Vol. 50, No. 12, p. 1066.
Luhn, H.P. 1959. "Keyword in Context Index for Technical Literature (KWIC Index)." Yorktown Heights, NY: IBM, Report RC 127. Also in: 1960. *American Documentation*, Vol. 11, pp. 288–295.
Parkins, P. V. 1963. "Approaches to vocabulary management in permuted title indexing of Biological Abstracts." In" *Automation and Scientific Communication Part 1, Proceedings of the 26th Annual Meeting of the American Documentation Institute*, pp. 27–28, Washington, DC: ADI.
Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. "Psychological Aspects of Natural Language Use: Our Words, Our Selves." Annual Review of Psychology, Vol. 54, pp. 547–77.
For an excellent socio-linguistic use of Pennebaker's work on function words, see:
Danescu-Niculescu-Mizil, Christian, Lilian Lee, Bo Pang, Jon Kleinberg. 2012. "Echoes of Power: Language Effects and Power Differences in Social Interaction." *Proc. 21st Int. Conf. World Wide Web*, Apr. 16–20, pp. 699–708. New York: Assoc. Comput. Mach.

Part XI (Weeks 14-15, April 12-14, April 19-21): A question of style

Back to the CoNLL table and what it reveals about style
Text readability: What grade level does a text require to be comprehensible?
Sentence complexity: Measuring and visualizing linguistic complexity
Analyzing vocabulary
N-grams and style
The use of function words, nominalization and passive forms as denial of agency
Using Gender Guesser for gender attribution: Who wrote this text?

Required readings:

Gender Guesser http://www.hackerfactor.com/GenderGuesser.php#About

Franzosi, Roberto, Gianluca De Fazio, and Stefania Vicari. 2012. "Ways of Measuring Agency and Action: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875-1930)." In: pp. 1-41, Tim Liao (ed.), *Sociological Methodology*, Vol. 42.

Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong. 2013. "From high heels to weed attics: a syntactic investigation of chick lit and literature." *Proceedings of the Second Workshop on Computational Linguistics for Literature*, pp. 72–81, Atlanta, Georgia, June 14, 2013.

Pennebaker, James W. and Laura A. King. 1999. "Linguistic Styles Language Use as an Individual Difference," *Journal of Personality and Social Psychology*, Vol. 77, No.6, pp. 1296-1312.

Suggested readings:

Pakhomov, Serguei, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. "Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing". *Behavior Research Methods*, Vol. 43, No. 1, pp. 136–144.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal*, Vol. 8, No. 1.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003a. "Gender, Genre, and Writing Style in Formal Written Texts," *Text*, Vol. 23, No. 3, pp. 321–346.

Kestemont, Mike. 2014. "Function Words in Authorship Attribution: From Black Magic to Theory?" *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 59–66, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Polio, Charlene and Hyung-Jo Yoon. 2018. "The reliability and validity of automated tools for examining variation in syntactic complexity across genres." *International Journal of Applied Linguistics*, Vol. 28, pp. 165-188.

Roark, Brian, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 7, pp. 2081-2090.

Tabata, Tomoji. 1995. "Narrative Style and the Frequencies of Very Common Words: A Corpus-Based Approach to Dickens's First Person and Third Person Narratives." *English Corpus Studies*, No. 2, pp. 91-109.

Frazier, Lyn 1985. "Syntactic Complexity." In: pp. 129-189, D. R. Dowty, L. Karttunen, and A. M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computation, and Theoretical Perspectives*. Cambridge: Cambridge University Press.

Yngve, Victor 1960. "A model and a hypothesis for language structure." *Proceedings of the American Philosophical Society*, Vol. 104, No. 5, pp. 444-466.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2013. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas". *Behavior Research Methods*, Vol. 46, pp. 904–911.

For a state-of-the-art review of authorship attribution, see

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. "Surveying Stylometry Techniques and Applications." *ACM Computing Surveys*, Vol. 50, No. 6, pp. 86:1–86:36, November.
Burrows, J.F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press; Oxford University Press.
Juola, Patrick. 2013. "Rowling and "Galbraith": An Authorial Analysis." URL http://languagelog.ldc.upenn.edu/nll/?p=5315.

Epilogue (Week 16, December 7): Digital humanities: A game changer?

On visual rhetoric

Required readings:

Franzosi, Roberto. 2015. "Of Stories and Beautiful Thing s: Digital Scholarship, Method, and the Nature of Evidence." Unpublished manuscript.
Healy, Kieran and James Moody. 2014. "Data Visualization in Sociology," *Annual Reviews of Sociology,* Vol. 4, pp. 105–28.

Suggested readings:

"Ad-writers are some of the most skilled rhetoricians in our society." (Edward P.J. Corbett and Robert J. Connors) Whatever else data visualization does… hopefully, it contributes to creating persuasive evidence. And if it is persuasive, it is rhetorical, rhetoric being the art of persuasion.

McQuarrie, Edward F. and David Glen Mick. 1996. "Figures of Rhetoric in Advertising Language." *The Journal of Consumer Research*, Vol. 22, No. 4, pp. 424-38.
Kostelnick, Charles. 2007. "The Visual Rhetoric of Data Displays: The Conundrum of Clarity," *IEEE Transactions on Professional Communications*, Vol. 50, No. 4, pp. 280–94.
Moretti, Franco. 1998 (1997). *Atlas of the European Novel, 1800-1900.* London: Verso.
Tufte, Edward R. 2006. *Beautiful Evidence.* Cheshire, CN: Graphics Press LLC.
Tukey, John W. 1969. "Analyzing Data: Sanctification or Detective Work?" *American Psychologist*, Vol. 24, No. 2, pp. 83-91.
Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory." *The American Statistician*, Vol. 34, No. 1, pp. 23-25.
Wainer, Howard. 1984. "How to Display Data Badly," *American Statistician*, Vol. 38, No. 2, pp. 137–47.
Gold, Matthew K. (ed.). 2012. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.
Tom, Gail and Anmarie Eves. 1999. "The Use of Rhetorical Devices in Advertising." *Journal of Advertising Research*, Vol. 39, July-August, pp. 39-43.
Forceville, Charles. 1996. *Pictorial Metaphor in Advertising*. London: Routledge.
Dyer, Gillian. 1988[1982]. "Chapter 8. The Rhetoric of Advertising", In: pp. 127-150, *Advertising as Communication*. Oxford: Routledge.
Leigh, James H. 1994. "The Use of Figures of Speech in Print Ad Headlines." *Journal of Advertising*, Vol. 23, No. 2, pp. 17-33.

McQuarrie, Edward F. and David Glen Mick. 1999. "Visual Rhetoric in Advertising: Text-Interpretive, Experimental, and Reader-Response Analyses." *The Journal of Consumer Research*, Vol. 26, No. 1 pp. 37-54.
Scott, Linda M. 1994. "Images in Advertising: The Need for a Theory of Visual Rhetoric." *The Journal of Consumer Research*, Vol. 21, No. 2, pp. 252-73.
Bush, Alan J. and Gregory W. Boller. 1991. "Rethinking the Role of Television Advertising during Health Crises: A Rhetorical Analysis of the Federal AIDS Campaigns." *Journal of Advertising*, Vol. 20, No. 1, pp. 28-37.
Barnard, Malcolm. 2005. "Metaphor/metonymy/synechdoche". In" pp. 50-54, *Graphic Design as Communication*. Abingdon, UK: Routledge.

Tufte has been a leading scholar on data visualization. Bertin, Cleveland, and Wilkinson are "classical" readings on data visualization. Some of the other readings, Yau in particular, represent the current state of the art on data visualization.

Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
Tufte, Edward R. 2003. *The Cognitive Style of PowerPoint*. Cheshire, CT: Graphics Press.
Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart.
Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart.
Bertin Jaques. 1967 (2010). *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands, CA: ESRI Press.
Wilkinson, Leland. 1995 (2005). *The Grammar of Graphics*. Second edition. New York: Springer.
Yau, Nathan. 2012. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.
Munzner, Tamara. 2014. *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.
Card, Stuart K., Jock D. Mackinlay, Ben Shneiderman (eds.). 1999. *Readings in Information Visualization: Using Vision to Think.* San Diego, CA: Academic Press.
Spence, Robert. 2014. *Information Visualization: An Introduction*. Third edition. New York: Springer.
Ware, Colin. 2012. *Information Visualization: Perception for Design.* Third edition. Waltham, MA: Elsevier.
Cleveland, William S. and Robert Mcgill. 1984. "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, Vol. 79, No. 388, pp. 807-22.
Funkhouser, H. Gray. 1937. "Historical Development of the Graphical Representation of Statistical Data," *Osiris*, Vol. 3, pp. 269-404.
Kosslyn, Stephen M. 1987. "Understanding Charts and Graphs." DTIC unpublished document.
McGill, Robert, John W. Tukey and Wayne A. Larsen. 1978. "Variations of Box Plots." *The American Statistician*, Vol. 32, No. 1, pp. 12–16.
Wickham, Hadley and Lisa Stryjewski. 2011. "40 Years of Boxplots." Unpublished manuscript.
Tufte, Edward R. 2001 [1983]. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
Anscombe, Francis J. 1973. "Graphs in statistical analysis." *American Statistician*, Vol. 27, pp. 17–21.
Friendly, Michael and Daniel Denis. 2005. "The Early Origins and Development of the Scatterplot." *Journal of the History of the Behavioral Sciences*, Vol. 41, No. 2, pp. 103–130.

Marshall, Alfred. 1885. "On the Graphic Method of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume (Jun. 22 - 24, 1885), pp. 251-260.

Keynes, John M. 1938. "Review of H.G. Funkhouser, Historical Development of the Graphical Representation of Statistical Data." *Economic Journal*, Vol. 48, No. 190, pp. 281–82.

Spence, Ian. 2005. "No Humble Pie: The Origins and Usage of a Statistical Chart," *Journal of Educational and Behavioral Statistics*, Vol. 30, No. 4, pp. 353–368.