Spring 2016 ***Big/Small Data & Visualization*** Soc 489/Ling 485

Emory University

Roberto Franzosi

Office Room No. 212 Tarbutton Hall
Email rfranzo@emory.edu

Lectures We 5:30PM - 8:30PM Tarbutton Hall 218
Office Hours MoWe 2:30-4:00 or by appointment (please, use email for contacts)

Alberto Purpura (alberto.purpura@studenti.unipd.it), MA student in Computer Science at the University of Padua (Italy), did much of the work to link PC-ACE to data visualization programs and NLP tools.

COURSE OBJECTIVES

The course deals with new tools of data analysis and visualization. Many of these tools have been developed in conjunction of new technologies of data mining and extraction from large text corpora made 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. The course will illustrate some of the cutting-edge projects in digital humanities such as David Eltis's *Trans-Atlantic Slave Trade* website (http://www.slavevoyages.org), Hank Klibanoff's *Georgia Civil Rights Cold Cases Project* (https://scholarblogs.emory.edu/emorycoldcases) or Roberto Franzosi's *Georgia Lynchings Project (1875-1940)*.

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. We will try to focus on freeware software, from *Gephi* to *Cytoscape*, *Palladio*, *Google Earth*, *QGIS*, *CartoDB*, *TimeMapper*, *Google Fusion Tables*.

Using Natural Language Processing (NLP) tools (based on the Stanford parser *CoreNLP*) the course will show how to analyze large corpora of text and how to visualize the information extracted, through *Excel* charts or *Wordle*, *Tagcrowd*, *Voyant* displays of word frequencies and network graphs of word co-occurrences. Other NLP tools and the visualization of the information they make available will also be introduced, from topic modeling (*Mallet*, *Stanford Topic Modeling Toolbox*) to Google *Word2Vec* (vectors representations of words, shown to capture many linguistic regularities of a corpus). The properties of such tools as Google ngrams and Bookworm will be explored. The peculiarities of dealing with words as data will be discussed (where and how can you get corpus data? How can you convert images to text? How can you check for spelling errors, differences in documents?).

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? If color, form, movement, in Kandinsky's view, are the distinctive weapons of art (and beauty), the new visualization techniques – all based on color, shape, and movement – are 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?

Learning outcomes

By the end of term, students are expected to be able:

1. Become familiar with NLP tools
2. To deal with large bodies of text ("corpus") with NLP tools
3. Visualize the NLP results
4. Become familiar with a large number of data visualization tools
5. To make public presentations before an audience
6. To write a research report

COURSE REQUIREMENTS

The course requires students:

1. to carry out **three sets of homework** based on NLP analyses and network and GIS representations of their corpus data;
2. to make individual or group **presentations on specific software** and edit/write appropriate **TIPS files**; you would get **extra points if you prepare screen recordings** on how PC-ACE connects to the software you are presenting (you could use http://screencast-o-matic.com/home for this purpose);
3. to make individual **presentations of research results** to the class at the end of term;
4. to write a **final paper** based on the analyses of their corpus data.

Students using an Apple laptop will be expected to install Windows on their laptop. There are a couple of approaches to this, via Apple Boot Camp and/or Parralels. Unfortunately, many of the applications we will use are Windows based. This is the only expense students will be required to sustain since all readings will be placed on reserve.

1. **Boot camp**: you can download it for **free** and install it at https://www.apple.com/support/bootcamp/ **For Boot camp to work your Mac must have an Intel processor. But Apple transitioned to Intel processors in the mid-2000s so you should be OK. But… please check your processor!**
2. **Parallels**: you can purchase a student license for **$39.95** (and volume pricing is also available for further discount) from the Emory website http://it.emory.edu/software/, then click on Academic Software (for Personal Purchases) which will redirect to

http://www.academicsuperstore.com/ Parallels at http://www.academicsuperstore.com/product/search?qk_srch=parallel&x=0&y=0

Both Parallels Desktop and Boot Camp allow you to run Windows on a Mac. One drawback of Boot Camp is that you have to reboot your Mac every time you want to switch between Mac and Windows. If this is not too often, it is a cheaper solution, and perhaps more reliable, than parallel.

Then, you need to **purchase** a copy of **Microsoft Windows** from the Academic Superstore with student discount at http://www.academicsuperstore.com/products/Microsoft/Windows+10/1605613 (USB Drive @ **$118.99**) (and volume pricing is also available) or from Amazon http://www.amazon.com/Microsoft-Windows-Home-Flash-Drive/dp/B01019T6O0/ref=sr_1_4?ie=UTF8&qid=1453734420&sr=8-4&keywords=windows+10 (Download @ **$109.99**).

You will also need a copy of **Microsoft Access, free from Emory**.

IT Services at Emory would recommend the following:

1. Visit Student Technology Support (STS) with your laptop for a preliminary consultation on what you will need for the installation (STS website: http://it.emory.edu/studentdigitallife/support/student-technology-support/; STS physical location: first floor of the Woodruff Library);
2. Ask STS personnel if it is okay to download the installer from online or if you need the USB drive installer;
3. Based on feedback from STS, purchase Windows 10 (Emory is now considering extending to students, and not just to staff, free Windows licenses; so, make sure to check this out before doing any purchases!);
4. Return to STS for Bootcamp setup and Windows and Access Installation;
5. **Allow plenty of time for STS to do the installation (at least one hour).**

Please, download the following freeware software that you will be using for textual analysis and data visualization:

1. **PC-ACE (Program for Computer-Assisted Coding of Events) the setup package will include the Stanford CoreNLP, Mallet, Word2Vec, KWIC and several data visualization options; PC-ACE only works under Windows operating system;**
2. **Stanford CoreNLP** http://nlp.stanford.edu/software/corenlp.shtml
3. **Mallet** http://mallet.cs.umass.edu/download.php
   (Stanford CoreNLP and Mallet and installed directly by PC-ACE)
4. **TACIT** http://tacit.usc.edu/download.html
5. **WordNet** https://wordnet.princeton.edu/wordnet/download/
6. **Gephi** http://gephi.org/users/download/
7. **Cytoscape** http://www.cytoscape.org/download.php
8. **Tableau** https://public.tableau.com/s/

9. **QGIS** https://www.qgis.org/en/site/forusers/download.html
10. **Google Earth Pro** http://www.google.com/earth/download/gep/agree.html
11. **Mondrian** http://www.theusrus.de/Mondrian/ (bottom of page, under downloads)
12. **OpenRefine (former Google Refine)** http://openrefine.org/download.html

In addition, we will be using the following web-based data visualization software:

1. **Voyant** http://voyant-tools.org/
2. **Google Fusion Tables** https://support.google.com/fusiontables/answer/2571232
3. **Raw** http://raw.densitydesign.org/
4. **TimeMapper** http://timemapper.okfnlabs.org/
5. **CartoDB** https://cartodb.com/
6. **Palladio** http://palladio.designhumanities.org/#/
7. **Google ngrams** https://books.google.com/ngrams
8. **Bookworm** http://bookworm.culturomics.org/
9. **Wordle** http://www.wordle.net/create
10. **TagCrowd** http://tagcrowd.com/
11. **Tagul** https://tagul.com/
12. **Taxedo** http://www.tagxedo.com/
13. **Wordclouds** http://www.wordclouds.com/

You should also download a set of texts of interest to you that you will want to analyze with NLP tools and visualize using the variety of data visualization tools illustrated in the course. These texts could be:

1. **tweets**
2. **blogs**
3. **newspaper articles**
4. **US Congress bills** (https://www.congress.gov/)
5. **US presidential speeches** (http://www.presidency.ucsb.edu/data.php)
6. **corporate/university mission statements**
7. **social science & history qualitative data**; see the US academic data depository of ICPSR of the University of Michigan (http://www.icpsr.umich.edu/index.html) or the British equivalent of the UK Data Service (https://www.ukdataservice.ac.uk/); the collection at Qualitative Data Repository (https://qdr.syr.edu/deposit), the Murray Research Archive at IQSS Harvard University* (http://murray.harvard.edu/dataverse)
8. **oral history archives**; see the list provided by the Oral History Association, (http://www.oralhistory.org/centers-and-collections/)
9. **transcribed in-depth interviews**
10. **social science journal abstracts** (http://ssrn.com/en/)
11. **NYT book reviews;** see the NYT API (http://developer.nytimes.com/docs/books_api/)
12. **song lyrics**; see, for example, the collection provided by AZlyrics (http://www.azlyrics.com/a/archive.html)
13. **books**; see the free collections at Open Library (https://openlibrary.org/) or at Hathi Trust Digital Library (https://www.hathitrust.org/); many older books are also available in Google Books (https://books.google.com/) and in other archives (e.g., The Gutenberg

Project https://www.gutenberg.org/, Internet Archive https://archive.org/ , The OAIster database http://www.oclc.org/oaister.en.html)

14. **diaries & autobiographies**
15. **letters (epistolary)**
16. **folktales (e.g., Afanasiev's collection of Russian folktales analyzed by Propp)**

If you do not have a corpus of interest, some will be made available to you, in particular:

1. Afanasiev's collection of fairy tales
2. Hundreds of newspaper articles on Georgia lynchings (1875-1940)

* On the Murray Research Archive see the paper: James, Jacquelyn B. and Annemette Sørensen. 2010. "Archiving Longitudinal Data for Future Research: Why Qualitative Data Add to a Study's Usefulness)," *FQS*, Vol. 1, No. 3, Art. 23 (http://www.qualitative-research.net/index.php/fqs/article/view/1040/2249#gref).

Web scraping. If you are obtaining your corpus from the web, you can copy and paste documents, perhaps from different websites. However, **web scraping** may provide a more efficient solution. Web scraping is the process of automatically collecting information from the World Wide Web through specialized software programs. A good, **freeware** option is **OutWit Hub**. While the full version of OutWith Hub costs around $89, the freeware option will probably serve you well. You can download it at http://www.outwit.com/products/hub/. Scraping requires knowledge of the data structure of each website where data are taken from. Scraping will be more efficient than human copy-and-paste if the documents to be scraped are stored under the same website (so that knowledge of only one type of data structure is required); otherwise, you may be better off by copying and pasting.

When you deal with digital material, you need different tools for combining files and converting files from different formats to a TXT format (all NLP tools deal with txt files only). PC-ACE has different routines to combine Word/txt files and convert Word files to TXT. But to convert pdf files to doc or txt you will need an external program. Use one of the many web-based tools, such as *RTF to PDF* (https://online2pdf.com/convert-rtf2pdf). In the conversion of a pdf file to txt, the file must not contain any images or the conversion will fail. The conversion will also fail if your pdf file is an image file. You will need, first, to convert the image to OCR (optical character reader). Acrobat Pro will do that for you. Alas, not Acrobat Reader and Acrobat Pro is expensive. If you do not have Acrobat Pro, since you will only have to do this once, just go to any of the computer labs on campus and use Acrobat Pro to convert your pdf image files.

COURSE PREREQUISITES

There are no formal prerequisites for the course, except for a general 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.

Deadlines and important dates

Term break:	March 7-9
Last day of classes:	April 25

Grading

Grading will be based on the following items:

1. *homework* (30%). Students are expected to hand in homework on the techniques presented during the previous weeks using the corpus data of their choice. A total of three pieces of homework are expected to be handed in, on the following topics: NLP, network graphs, GIS maps. If the corpus has no geographic data, for the GIS assignment, students are expected to use a different corpus.
2. *participation* (10%). Students are expected to attend classes regularly (attendance is enforced through a sign-up sheet) and contribute to class discussion.
3. *presentations (20%).* Students are expected to make two types of presentations to the class:
   a. *software presentations and TIPS files* (10%). Students are expected to make a presentation of one or more specific software (e.g., Palladio) illustrating to the class the software capabilities. For the software students present, they are expected to make sure that the available TIPS files are correct or, if no TIPS files are available, they are expected to write one or more, as appropriate. **What is a TIPS** file? A TIPS file is a document originally written for PC-ACE users that provides help on a specific issue (e.g., on the use of Gephi). It is meant as basic, first-time-user help on what users could do in a software they do not know.
   b. *presentation of research results (10%).* At the end of term students will present the overall results of the analyses of their corpus data (description of the corpus, NLP and visualizations)
4. *final research paper* (40%). Students are expected to write a final paper where they bring together all the analyses they have performed on the text corpus. Students are welcome to organize their paper in the standard format – Introduction, Literature Review, Data & Methods, Empirical Results, Conclusions, Bibliography – but they are also encouraged to experiment with creative writing (provided that all relevant information of the standard format is still provided). **Students should aim to write a publishable quality paper. The paper should include plots, charts, graphs, and links to dynamic visualizations. The paper should be 6,000 words in length excluding visuals.**

Students who are not satisfied with a grade received arc 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://www.sph.emory.edu/cms/current_students/enrollment_services/honor_code.html

COURSE OUTLINE

Part I: Visualizing data

Week 1: Is a picture worth a thousand words?
Week 2: Traditional data visualizations: boxplots, time plots, scatterplots, pie charts and bar charts

Part II: Analyzing and visualizing big data: Things to do with words

Week 3: Big data and NLP (Natural Language Processing)
Week 4: Stanford CoreNLP and the CoNLL table: Things to do with CoNLL table data
Week 5: Word co-occurrences (KWIC) & Word2Vec
Week 6: Topic modeling & Semantic dictionaries (WordNet)
Week 7: Visualizing words: ngrams, Word clouds (Pictures worth billions of words)
Week 8: Quantitative Narrative Analysis (QNA): How does it work and what can you do with the 5 Ws + H of narrative?

Week 9: Spring Break!!!

Part III: Networks, time and space

Week 10: Visualizing networks
Week 11: Visualizing time and space
Week 12: Visualizing maps

Part IV: What have we learned?

Week 13: A game changer? Digital humanities (or Digital scholarship), beautiful evidence, and visual rhetoric
Week 14: Project presentations
Week 15: Project presentations

Readings

Readings for the course come from books and journal articles or book chapters. All reading material has been placed on **Ereserve** and physical copies of most of the required books are on Reserve:

1. Tufte, Edward R. 2006. *Beautiful Evidence.* Cheshire, CN: Graphics Press LLC.
2. Tufte, Edward R. 2001 [1983]. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
3. Tufte, Edward R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
4. Tufte, Edward R. 2003. *The Cognitive Style of PowerPoint*. Cheshire, CT: Graphics Press.
5. Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart.
6. Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart.

7. Bertin Jaques. 1967 (2010). *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands, CA: ESRI Press.
8. Wilkinson, Leland. 1995 (2005). *The Grammar of Graphics*. Second edition. New York: Springer.
9. Yau, Nathan. 2012. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Indianapolis, IN: Wiley.
10. Munzner, Tamara. 2014. *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.
11. Card, Stuart K., Jock D. Mackinlay, Ben Shneiderman (eds.). 1999. *Readings in Information Visualization: Using Vision to Think.* San Diego, CA: Academic Press.
12. Spence, Robert. 2014. *Information Visualization: An Introduction*. Third edition. New York: Springer.
13. Ware, Colin. 2012. *Information Visualization: Perception for Design.* Third edition. Waltham, MA: Elsevier.
14. Moretti, Franco. 2013. *Distant Reading*. London: Verso.
15. Moretti, Franco. 2005. *Graphs, Maps, Trees. Abstract Models for a Literary History*. London: Verso.
16. Moretti, Franco. 1998 (1997). *Atlas of the European Novel, 1800-1900*. London: Verso.
17. Forceville, Charles. 1996. *Pictorial Metaphor in Advertising*. London: Routledge.

There are some weeks with very heavy readings. And books are long... Read enough to know what they are saying. Some weeks are heavier in readings than others. Try to distribute your workload appropriately. I have separated readings in the standard format of Required and Suggested readings. However, I have placed comments in red under the Suggested readings labels. You are strongly encouraged to take at least a quick look to those readings to familiarize yourself with the basic language and arguments on specific topics.

Part I: Visualizing data

January 13
Week 1: Is a picture worth a thousand words?

Required readings:

Healy, Kieran and James Moody. 2014. "Data Visualization in Sociology," *Annual Reviews of Sociology,* Vol. 4, pp. 105–28.
Wainer, Howard. 1984. "How to Display Data Badly," *American Statistician*, Vol. 38, No. 2, pp. 137–47.
Tufte, Edward R. 2001 [1983]. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Suggested readings:

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.

January 20
Week 2: Traditional data visualizations: boxplots, time plots, scatterplots, pie charts and bar charts

Software: Microsoft Excel

Required readings:

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.

Suggested readings:

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.

Part II: Analyzing and visualizing big data: Things to do with words

January 27
Week 3: Big data and NLP (Natural Language Processing)

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. Talk by Nello Cristianini on Big Data ("Patterns in Media Content) https://www.youtube.com/watch?v=mmWRNRPb0W0
Quick video. Nello Cristianini's visualization of millions of tweets ("Mood Changes of UK during 2009-2012") https://www.youtube.com/watch?v=gG5ZH2JfqIU
Quick video. Nello Cristianini's visualization of millions of newspaper articles ("Key Actors in US Presidential Elections 2012 – primaries") https://www.youtube.com/watch?v=ptH5FkKSvvU
Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Kirschenbaum, Matthew G. 2009. "The Remaking of Reading: Data Mining and the Digital Humanities." Talk given at the 2009 National Science Foundation Symposium on the Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation.

Suggested readings:

Take a quick look at some of these readings, especially Mohr et al. Familiarize yourself with what the ready availability of digital newspaper archives would allow you to do/and how (see Snowsill et al., Zervanou ett al. Seguin et al.).

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

Lansdall-Welfare, Thomas, Saatviga Sudhahar, Giuseppe A. Veltri, and Nello Cristianini. 2014. "On the Coverage of Science in the Media: A Big Data Study on the Impact of the Fukushima Disaster," IEEE International Conference on Big Data (Big Data). Washington DC, 27-30 Oct. 2014.

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.

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

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.

Seguin, Charles. 2015 web download. "Scraping Historical Newspaper Archives: The Transformation of Public Lynching Discourse in the US." http://badhessian.org/2014/01/scraping-historical-newspaper-archives-the-transformation-of-public-lynching-discourse-in-the-us/

February 3
Week 4: Stanford CoreNLP and the CoNLL table: Things to do with CoNLL table data

Software: Stanford CoreNLP

Required readings:

Franzosi, Roberto. NLP TIPS files.

February 10
Week 5: Word co-occurrences (KWIC) & Word2Vec

Software: PC-ACE & Word2Vec (from PC-ACE)

Required readings:

Franzosi, Roberto. NLP TIPS files.

February 17
Week 6: Topic modeling & Semantic dictionaries (WordNet)

Software: Mallet, WordNet

For the adventurous; **mandatory for Computer Science major/minor**.

Try your hand at:

1. **Apache OpenNLP** (https://opennlp.apache.org/). From their website we read: The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning.

2. **NLTK (Natural Language ToolKit)** (http://www.nltk.org/index.html)**.** From their website we read: NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries

Required readings:

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

Some words of caution on the big-data revolution… Video. Talk by Nello Cristianini, "The Big-Data Revolution and its Impact on Science and Society." https://www.youtube.com/watch?v=PzicexAmycA

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.

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.
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.
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.
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.
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.

February 24
Week 7: Visualizing words: ngrams and Word clouds (Pictures worth billions of words)

Software: Google ngrams, Bookworm, Wordle, TagCrowd, Tagul and Taxedo (Tagul and Taxedo allow to draw word clouds in specific shapes)

Required readings:

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

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
Aiden, Erez Lieberman and Jean-Baptiste Michel. 2011. "A picture is worth 500 billion words". http://tedxtalks.ted.com/video/TEDxBoston-Erez-Lieberman-Aid-2
Michel, Jean-Baptiste et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 14 January 2011, Vol. 331, pp. 176-182.
Nunberg, Geoffrey. 2009. "Google's Book Search: A Disaster for Scholars." *The Chronicle of Higher Education,* August 31, 2009.
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.

Suggested readings:

Become familiar with the basic language of culturomics!

Gold, Matthew K. (ed.). 2012. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.
Schwartz, Tim. 2011. "Culturomics Periodicals Gauge Culture's Pulse." *Science,* Vol. 332, 1 April 2011, p. 35-36.

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).

March 2

Week 8: Quantitative Narrative Analysis (QNA): How does it work and what can you do with the 5 Ws + H of narrative?

Software: PC-ACE

Required readings:

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

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.

Suggested readings:

Franzosi, Roberto. 2014. "Analytical Sociology and Quantitative Narrative Analysis: Explaining Lynchings in Georgia (1875–1930)." In: pp. 127-148, Gianluca Manzo (ed.), *Analytical Sociology: Actions and Networks*. Hoboken, NJ: Wiley.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis (Quantitative Applications in the Social Sciences)*. Beverly Hills, CA: Sage.

Automating QNA:

Required readings:

Sudhahar, Saatviga, Roberto Franzosi, Nello Cristianini. 2011. "Automating Quantitative Narrative Analysis of News Data," *Journal of Machine Learning Research*, Vol. 17, pp. 63-71.

Suggested readings:

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!

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

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

Lendvai1, 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).

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.

March 9

Week 9: Spring Break!!!

Part III: Networks, time and space

March 16

Week 10: Visualizing networks

Software: Cytoscape, Gephi, Google Fusion Tables, Raw, NodeXL, Palladio

For the adventurous; **mandatory for Computer Science major/minor.**

Try your hand at **D3.JS** (http://d3js.org/). From their website we read: D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Raw http://raw.densitydesign.org/

Required readings:

Franzosi, Roberto. Networks TIPS files.

Moody, James, Daniel McFarland and Skye Bender-deMoll. 2005. "Dynamic Network Visualization: Methods for Meaning with Longitudinal Network Movies," *American Journal of Sociology*, Vol. 110, No. 4, pp. 1206–41.

March 23
Week 11: Visualizing time and space

Software: Google Earth Pro, GeoNames, OpenStreetMap, TimeMapper

Required readings:

Franzosi, Roberto. Geocoding TIPS files.

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.
Rosenberg, Daniel and Anthony Grafton. 2010. *Cartographies of Time*. New York, Princeton Architectural Press.
Massey, Doreen. 2005. *For Space.* Thousand Oaks, CA: Sage.

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

March 30
Week 12: Visualizing maps

Software: CartoDB, Google Earth, Google Earth Pro, Mondrian, QGIS, Tableau, TimeMapper

Required readings:

Franzosi, Roberto. GIS TIPS files.

Part IV: What have we learned?

April 6
Week 13: A game changer? Digital humanities (or Digital scholarship), beautiful evidence, and visual rhetoric

Required readings:

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

On digital humanities and beautiful evidence:

Required readings:

Franzosi, Roberto. 2015. "Of Stories and Beautiful Things: Digital Scholarship, Method, and the Nature of Evidence." Unpublished manuscript.

Tufte, Edward R. 2006. *Beautiful Evidence.* Cheshire, CN: Graphics Press LLC.

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.

Kirschenbaum, Matthew G. 2011. "Digital Humanities Archive Fever." Plenary lecture at the Digital Humanities Summer Institute at the University of Victoria, June 2011. August 22, 2011 at 9:56 PM · https://vimeo.com/28006483

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.

Jockers, Matthew L. and David Mimno. 2013. "Significant Themes in 19th-Century Literature," *Poetics*, Vol. 41, No. 6, pp. 750-769.

Pannapacker, William. 2009. "The MLA and the Digital Humanities." *The Chronicle of Higher Education*, December 28, 2009.

Moretti, Franco. 2005. *Graphs, Maps, Trees. Abstract Models for a Literary History.* London: Verso.

Suggested readings:

Moretti, Franco. 1998 (1997). *Atlas of the European Novel, 1800-1900.* London: Verso.

On visual rhetoric:

Required readings:

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.
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.

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.

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.

April 13
Week 14

Project presentations. No readings.

April 20
Week 15

Project presentations. No readings.