“[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 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 2:30-3:45, Callaway Center C101 after the online period

Undergraduate TAs Eden Medina  
Email eden.medina@emory.edu  
Personal meeting room (Office hours by appointment)  
https://us02web.zoom.us/j/83947900879

Allison Reinhardt  
allison.brooke.reinhardt@emory.edu  
Personal meeting room (Office hours by appointment)  
https://emory.zoom.us/j/4339176194

DEADLINES AND IMPORTANT DATES

First day of class January 11  
Last day of class April 25  
Spring break March 7-11 no classes  
Homework Due each Sunday at midnight  
Software Problems Raise issue on GitHub  
Presentations A set of 3 presentations per team, due each Thursday for selected teams, starting on week 3

Table of Contents
COURSE OBJECTIVES

Tools of analysis and visualization of text data
Learning the language of Natural Language Processing (NLP)
Big data/small data
Visualization and a world of beauty. A game changer?

Why should YOU take this course: Learning outcomes

Welcome to the 21st century!
Learning outcomes
Ongoing measurement of learning outcomes

Is this a course for YOU?

No prerequisites
No prerequisites but... A hard course?
GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders, Videos
All you need to do is press buttons but... interpret results!
Download and install the NLP Suite and other software
NLP Suite welcome GUI
You need a work partner
You also need a corpus
What is a corpus?
What types of files do you need for your corpus (csv, pdf, docx)? Txt!!!
And where would you get this corpus?
Option 1: Work on corpora provided in the course
Option 2: Get your own corpus
Web scraping
Weekly homework assignments
Homework rubrics

GRADING
Participation (5%)
Presentations (35%) – Starting on week 3
Homework (60%)
Bonus points

HONOR CODE

WEEKLY HOMEWORK

WEEKLY TOPICS & READINGS

Required & suggested readings
Where will you find the readings?
Introduction (Week 1, January 11-13)
Big data and “distant reading”
Digital humanities: What is it?
File types doc, docx, rtf, txt, pdf) and what to do about it
NLP: What is it?
Preparing your corpus for “distant reading”
Becoming familiar with the suite of Java and Python NLP tools
Homework 1 (due Sunday January 16, at midnight)
Installing NLP software for distant reading
Part I (Week 2, January 18-20): Corpus Statistics and Words Visualization
Corpus Statistics
Visualization in Digital humanities
Word clouds
Excel charts (with hover-over effects)
Network graphs: Mapping relations
Knowledge graphs (KG) and HTML annotated files
Maps: Space (and time)
Homework 2 (due Sunday January 23, at midnight)
A basic look at a corpus via distant reading
Part II (Week 3, January 25-27): Topic Modeling & Word2Vec
What are the topics in your corpus?
Topic modeling via Gensim and MALLET
Word2Vec
Homework 3 CANCELLED (NOT due Sunday January 30, at midnight)
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
Homework 4 (due Sunday February 6, at midnight)
Topic modeling & Word2Vec
Part IV (Week 5, February 8-10): The CoNLL table, Named Entity Recognition (NER) and CoreNLP annotators
A closer look at the CoNLL table: Meet the NER, POSTAG, DEPREL tags
Searching the CoNLL table
Stanford CoreNLP annotators
Is there dialogue?
Are there people and organizations and differences in gender distribution?
Are there geographical locations?
Are there times?
Using WordNet: Does nature appear?
Using WordNet: Do nouns and verbs cluster in specific classes?
Homework 5 (due Sunday February 13, at midnight)
Parsers and CoNLL table
Part V (Week 6, February 15-17): From text to maps
Using CoNLL NER information to map locations
Geocoding
Visualizing time and space
Homework 6 (due Sunday February 20, at midnight)
NER location and GIS maps
Part VI (Week 7, February 22-24): Narrative and the 5 Ws
SVO Extraction & Visualization
Stanford CoreNLP enhanced dependencies parser
SENNA
Stanford CoreNLP OpenIE
Homework 7 (due Sunday February 27, at midnight) .......................................................... 30
Subject-Verb-Object (SVO) extractors .................................................................................. 30
Part VII (Weeks 8-9, March 1-3, March 8-10): Word N-grams and co-occurrences .......... 30
CoNLL table analyzer .......................................................................................................... 30
N-grams: What are they and what are they good for? ......................................................... 30
Google N-grams Viewer and Culturomics ......................................................................... 30
N-grams searches in the NLP Suite ..................................................................................... 30
Word co-occurrences searches ............................................................................................ 30
Single words/collocations searches .................................................................................... 30
Homework 8 (due Sunday March 6, at midnight) ............................................................... 32
Searching a corpus: CoNLL table, N-grams, co-occurrences, culturomics ...................... 32
SPRING BREAK March 8-10 no classes .............................................................................. 32
Homework 9 – NO homework due Sunday March 13 at midnight – Spring break .......... 32
Part VIII (Week 10, March 15-17): Knowledge-graphs/Knowledge-base systems (DBpedia
and YAGO) .......................................................................................................................... 32
DBpedia ............................................................................................................................... 32
YAGO .................................................................................................................................. 32
Dictionary-based annotation ............................................................................................... 32
html files .............................................................................................................................. 32
Homework 10 (due Sunday March 20, at midnight) ........................................................... 33
Knowledge-graphs/Knowledge-base systems (DBpedia and YAGO) ............................... 33
Part IX (Weeks 11-12, March 22-24, March 29-31): The world of emotions .................... 33
Sentiment Analysis: Capturing the feelings conveyed in the writing .................................. 34
WordNet ............................................................................................................................... 34
YAGO .................................................................................................................................. 34
ANEW ................................................................................................................................. 34
Hedonometer ........................................................................................................................ 34
SentiWordNet ....................................................................................................................... 34
Stanford CoreNLP sentiment analysis annotator ................................................................. 34
VADER ................................................................................................................................. 34
Homework 11 (due Sunday March 27, at midnight) ........................................................... 35
Sentiment analysis ............................................................................................................... 35
The “shape” of stories ........................................................................................................... 35
Data reduction algorithms: Hierarchical Clustering (HC), Singular Value Decomposition
(SVD), Non-Negative Matrix Factorization (NMF) ............................................................ 35
Homework 12 (due Sunday April 3, at midnight) ............................................................... 35
The shape of stories ............................................................................................................. 36
Part X (Week 13 April 5-7): Dissecting your corpus via the CoNLL table ......................... 36
Searching the CoNLL table for relationships between words ............................................ 37
Noun density and noun types .............................................................................................. 37
Verb modality: Ability, possibility, permission, and obligation .......................................... 37
Verb tense: past, future, gerundive ...................................................................................... 37
Verb voice: Active and passive verb forms ........................................................................ 37
Function words ("junk" words or "stop" words): pronouns, prepositions, articles, conjunctions, and auxiliary verbs ................................................................. 37
Pronouns and Coreference resolution ........................................................................................................ 37
The use of function words, nominalization and passive forms as denial of agency ............................... 37
Homework 13 (due Sunday April 10, at midnight) ................................................................................ 38
   Zooming into the CoNLL table ........................................................................................................ 38

Part XI (Weeks 14-15, April 12-14, April 19-21): A question of style ............................................... 38
   Back to the CoNLL table and what it reveals about style ............................................................... 38
   Text readability: What grade level does a text require to be comprehensible? .......................... 38
   Sentence complexity: Measuring and visualizing linguistic complexity ..................................... 38
   Analyzing vocabulary .................................................................................................................. 38
   N-grams and style ......................................................................................................................... 38
   Using Gender Guesser for gender attribution: Who wrote this text? ......................................... 38
Homework 14 (due Sunday April 17, at midnight) ........................................................................... 40
   A question of style ....................................................................................................................... 40

Epilogue (Week 15, April 19-21): Digital humanities: A game changer? ........................................... 40
   Homework 15 (due Sunday April 24, at midnight) ...................................................................... 42
   NLP tools and your corpus: The most significant findings .......................................................... 42
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

446WR fulfills the writing requirement

As the course deals mostly with automatic processing of texts, the issue of writing and style are implicitly at the core of the course: which verb voices are used (active or passive), which level of sentence complexity (as measured by different indices of sentence complexity), which semantic roles (e.g., agent and patient, experiencer, benefactor and beneficiary, messenger and receiver), which attributes (e.g., adjectives or adverbs) in conjunction with different nouns and verbs, which sentiments are expressed in sentences (negative, neutral or positive). Teaching writing is then a fundamental part of analyzing writing. The pros and cons of pure automatic analyses of texts (“distant reading” through a computer) are constantly brought up, with an emphasis of a constant dialogue between distant reading and close reading.

So… if a poetry course scares you… maybe this is a good option for fulfilling the writing
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 homework and homework rubrics.

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!

No prerequisites but... A hard course?

Perhaps. But not because there are impossibly hard homework or readings (some of them may be hard; but if you are not a computer scientist getting the gist of them is good enough). The course is demanding because there are readings and homework every week; and in order to fulfil the writing requirement (the W in Soc/Ling/QSS 446WR) you need to write at least 5 pages every week. But the rest is easy.
GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders, Videos

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 but… interpret results!

If you know how to do that, you are halfway there…

![The introductory Graphical User Interface (GUI) to the NLP Suite](image)

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.

<table>
<thead>
<tr>
<th>What is a parser?</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free/open-source parsers</td>
<td>1</td>
</tr>
<tr>
<td>The Stanford CoreNLP parsers</td>
<td>2</td>
</tr>
<tr>
<td>System requirements</td>
<td>2</td>
</tr>
<tr>
<td>Java</td>
<td>2</td>
</tr>
<tr>
<td>Input</td>
<td>2</td>
</tr>
<tr>
<td>Output: The CoNLL table</td>
<td>2</td>
</tr>
<tr>
<td>The named-entity recognition dependency parser output tags</td>
<td>3</td>
</tr>
<tr>
<td>Faulty results</td>
<td>3</td>
</tr>
</tbody>
</table>

Download and install the NLP Suite and other software

From GitHub ([https://github.com/NLP-Suite/NLP-Suite/wiki](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 2.**

1. **graduate students** can choose to work alone and on their own data if they prefer;
2. **undergraduate students** need instructor’s permission to work alone, only granted under special circumstances.

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

**Starting week 2, each team will only submit one homework for both team members. For the first week each student will have to submit an individual homework.**
You also need a corpus

What is a corpus?

A corpus is a larg(ish) collection of TEXT documents, the more the better (at least 100 documents), e.g., newspaper articles, blogs, short stories, or whatever. This text corpus will be the basis of your weekly analyses using different NLP tools and of your final paper.

What types of files do you need for your corpus (csv, pdf, docx)?Txt!!!

These texts should be in txt format (not doc, pdf, or other since NLP tools only work with txt formats). The NLP Suite has a set of functions to convert docx and pdf documents to txt.

Undergraduate students can choose to work on specific corpora and with specific partners of their choice:

1. Students without a selected partner and corpus will be randomly assigned corpus and partner after the add/drop period
2. For the first two weeks during the add/drop period, students can choose to work on any corpus for the assignments and/or alone and get full credit for their work.

And where would you get this corpus?

Option 1: Work on corpora provided in the course

We have several text corpora that you can analyze. The analyses of some of these corpora may lead to co-authored journal publications.

1. Gay men project
   376 personal narratives from gay men from 37 different countries
2. The Harry Potter books
   J. K. Rowling’s collection of 7 Harry Potter books
3. 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)
   Some 1300 NYT best-selling book reviews
5. Folktales
   A collection of hundreds of cross-national English, German, Chinese, Arabic, and Indian folktales

Option 2: Get your own corpus

1. blogs
2. newspaper articles
3. US Congress bills (https://www.congress.gov/); for an easier approach, see https://www.congress.gov/search?q=%22source%22:%22legislation%22)&searchResultViewType=expanded
4. corporate/university mission statements
5. 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)
6. oral history archives; see the list provided by the Oral History Association, (http://www.oralhistory.org/centers-and-collections/)
7. transcribed in-depth interviews
8. social science journal abstracts (http://ssrn.com/en/)
9. song lyrics; see, for example, the collection provided by AZlyrics (http://www.azlyrics.com/a/archive.html)
10. 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)
11. diaries & autobiographies
12. letters (epistolary)

NLTK, for those who know Python, has a great way for accessing various resources: https://www.nltk.org/book/ch02.html

Make sure you check the data in your corpus.
1. To repeat... you can only use txt-formatted utf-8 files (NLP tools only work with txt files in input).
2. Remove tables of contents, indices, weirdly formatted footnotes/endnotes, headers/footers, tables and figures. This material is not handled correctly by NLP tools.

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.
1. 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/. Another good freeware option is HTTrack (https://www.httrack.com/). 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.
2. If you are a **Python** programmer, you can also use the **BeautifulSoup** package (https://www.crummy.com/software/BeautifulSoup/bs4/doc/).

3. If you are an **R user** you can use the **rvest** package.

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). To convert pdf files to doc or txt you will need an external program. The NLP suite of stools that we use has a good Python conversion routine. You can also use one of the many web-based tools, such as **RTF to PDF** (https://online2pdf.com/convert-rtf2pdf). If your pdf file is an image file, you may 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.

**Weekly homework assignments**

The **weekly assignments**, 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.**

**Homework rubrics**

Each assignment is graded (0-100) and comments are provided. **Weekly rubrics for the homework are also provided**, detailing the scale for different points. **Every week, you will know exactly what you missed!** The standards of writing are repeatedly explained in class and stressed in the comments given to students. **Rubrics are posted under Files on CANVAS but also posted weekly 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**

This is an intensive computer and writing course.

Grading will be based on the following items:
**Participation (5%).** You are expected to attend classes regularly (attendance is enforced through a sign-up sheet) and contribute to discussion.

**Presentations (35%) – Starting on week 3** students or teams of students will make in-class presentation of their work. **3 presentations total will be scheduled.** 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, Wrd2Vec, 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.

**Homework (60%).** You are expected to carry out weekly homework that you will upload to CANVAS. Homework assignments will involve the use of specific NLP tools applied to specific corpus data (e.g., Stanford CoreNLP, Gensim, Mallet, sentence length visualization). You will need to present screenshots of your work and, especially, interpret your results with extensive writeups. You need to answer questions such as: what does the tool allow you to do? How does it work? What are its pros and cons? How do you interpret the results? What does each tool tell you about your data? How has the tool been developed/used in a scholarly community? Each homework will be graded out of 100 points. Make sure to include:

- screenshots of your work;
- engaged references to the readings.

Expect homework to take 5 or 6 hours in a combination of computer work and writing.

The homework reports should be at least 5 pages in length, including visuals.

Late homework will be automatically penalized by subtracting 10 points, unless prior permission was granted.

Homework will be graded broadly (but not strictly) following the rubric and returned within a week of due date.

Each homework must include at the top a statement with the % contribution of each partner.

**Bonus points.** Students with a good programming background can get extra points by carrying out specialized programming tasks to develop specific tools. But if you are not a programmer, you 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.** More demanding programming work can also be carried out instead of some weekly homework.
Attendance to synchronous class is mandatory and enforced by checking Zoom presence. Students who cannot attend on a regular basis should contact the professor or the TA. During the online period, every class session will be recorded and made available under special circumstances. Recorded class session are strictly private and cannot be made available outside the class.

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 HOMEWORK

In homework, please, provide screenshots and extensive write-ups of your findings. Homework submitted without screenshots will receive a ZERO grade. Homework writeup MUST engage extensively with the appropriate scholarly literature. Late homework will be automatically penalized by subtracting 10 points, unless prior permission was granted. Homework will be graded broadly (but not strictly) following the rubric and returned within a week of due date.

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?
File types doc, docx, rtf, txt, pdf) and what to do about it
NLP: What is it?
Preparing your corpus for “distant reading”
Becoming familiar with the suite of Java and Python NLP tools

Required readings:

NLP Suite GitHub wiki pages https://github.com/NLP-Suite/NLP-Suite/wiki
TIPS_NLP_Things to do with words NLP approach.pdf

https://machinelearningmastery.com/natural-language-processing/


https://youtube.com/seeapattern

Suggested readings:


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

The Digital Scholarship Lab at the University of Richmond, [http://dsl.richmond.edu/](http://dsl.richmond.edu/)
The Yale photographic site [http://photogrammar.yale.edu/](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](http://atlas.lib.uiowa.edu)

**Homework 1 (due Sunday January 16, at midnight)**

*Installing NLP software for distant reading*

1. Provide screenshots of successful installation of software on your computer.
2. What do those batch files in setup_Mac or setup_Windows do?
3. 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?
4. 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?

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

**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

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


Suggested readings:


Homework 2 (due Sunday January 23, at midnight)

A basic look at a corpus via distant reading

1. What do all the scripts in the Data & File Handling Tools and Pre-Processing Tools do?

Can you think of reasons why you would need them?

2. What is a corpus (Latin plural corpora)? If you chose to work on your own corpus, provide a one-page description of the corpus, detailing the reasons for selecting the corpus and the hunches about what to expect from an analysis of the corpus.

3. What is NLP? What are “distant reading” and “digital humanities.” Why distant? What are the pros and cons of distant reading?

4. What file types do NLP tools deal with (pdf, rtf, docx, txt)? What can you do if you have pdf files? What does it mean to have utf-8 compliant files and why is this an issue? Why can apostrophes and quote symbols give you problems? What can you do about it? Why does my csv file output look so messy?

5. Using the What’s in your corpus tool, get the basic statistics of your corpus (e.g., number of documents, sentences, words, n-grams)? What do they say?

6. Using word clouds programs (e.g., Tagxedo, Tagul, Python WordCloud), display the words of your corpus in the various programs. What do these NLP tools applied to your corpus tell you?

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

Required readings:

Franzosi, Roberto. NLP TIPS files.


For an interesting paper on topic modeling based on Gensim and with various practical recommendations and references, see:


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.


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

**Homework 3 CANCELLED (NOT due Sunday January 30, at midnight)**

**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.predictivatanalyticstoday.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=mmWRNRp0W0
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.


Homework 4 (due Sunday February 6, at midnight)

Topic modeling & Word2Vec

What topics does your corpus cover? Does topic modeling have an answer to that question for your corpus? Do Gensim and Mallet correctly categorize your corpus? Which tool performs better? In Gensim, what is the “ideal” distribution of topics in the Intertopic Distance Map (via multidimensional scaling)? In Gensim, what is the effect of varying the relevance metric λ and how do you interpret the results? What is Word2Vec?
Which words come together in a semantic space when you run Word2Vec on your corpus? What insights do you gain on your corpus from topic modeling and Word2Vec?

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

A closer look at the CoNLL table: Meet the NER, POSTAG, DEPREL tags
Searching the CoNLL table
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 other 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.

Zooming in

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.

Homework 5 (due Sunday February 13, at midnight)

CoreNLP parsers and annotators

Run the Stanford CoreNLP on your corpus to produce the CoNLL table and provide a two-page description of your results with separate screenshots of results. Do not merge the files in your corpus before parsing. The parser will identify the different files in the CoNLL table. What does the CoNLL table tell you? What are the various fields? Make sure to define the terms Form, lemma, POSTAG, DEPREL, NER. Use the CoNLL table analyzer to address “meaningful” questions about significant words and word relations in your corpus (e.g., which adjectives are used for which nouns). What significant questions about your corpus do these NLP tools allow you to answer? What do the various CoreNLP annotators tell you about your corpus?

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.


Suggested readings:


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

Homework 6 (due Sunday February 20, at midnight)
NER location and GIS maps
Using the NER information of your corpus (if your corpus contains locations; use Professor Dumbass story in the sampleData subfolder of NLP if there are no locations in your corpus), extract location information, geocode locations and map them using Google Earth Pro and Google Maps for heatmaps using the script GIS_main.py. What is the difference between the two types of maps? What kind of information do you need to draw dynamic GIS maps? How can you make your maps more beautiful, more vivid, following geographer Peirce Lewis’s recommendations (1985)?

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
SEenna
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.


Suggested readings:


(Found under Murchú or zip will not zip)


Homework 7 (due Sunday February 27, at midnight)

**Subject-Verb-Object (SVO) extractors**

Using the SVO extractor tool (SVO_main.py), analyze your corpus to extract the Who, the What, the When and Where of narrative. Do three algorithms available in the NLP Suite for SVO extraction provide similar/different results? In which way? Run the SVO tools with and without coreference resolution (via Stanford CoreNLP). What difference does it make? Finally, if you run the coreference resolution of The Tree Little Pigs story that you will find in the NLP\lib\sampleData subdirectory how does CoreNLP coref do? Can you use the manual coref resolution GUI to resolve those cases not dealt with (or poorly dealt with) automatically?

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

**Week 8: March 1-3**

*CoNLL table analyzer*

*N-grams: What are they and what are they good for?*

*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](https://www.youtube.com/watch?v=WtJ50v7qByE&t=19s)


Suggested readings:


**Homework 8 (due Sunday March 6, at midnight)**

*Searching a corpus: CoNLL table, N-grams, co-occurrences, culturomics*

Write a five-page report on the results of the SEARCH TOOLS encountered this week and applied to your corpus (N-Grams and Co-Occurrences Viewer). Do these tools give you more mileage than the CoNLL table analyzer searches? Make sure to define such terms as N-Grams, and word co-occurrences and, again, to address “meaningful” questions about significant words and word relations in your corpus. What are the differences between Google Ngram Viewer and the Java tool in the NLP Suite? Why would you want to duplicate routines? What does culturomics mean and what are the pros and cons of culturomics? Can Big Data and the hype of culturomics lead to the end of theory?

**Week 9: March 8-10**

**SPRING BREAK March 8-10 no classes**

**Homework 9 – NO homework due Sunday March 13 at midnight – Spring break**

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

*DBpedia*

*YAGO*

*Dictionary-based annotation*

*html files*
Required readings:

Franzosi, Roberto. NLP TIPS files.


Suggested readings:


Homework 10 (due Sunday March 20, at midnight)
Knowledge-graphs/Knowledge-base systems (DBpedia and YAGO)

Using the DBpedia, YAGO, and dictionary script, how can you use the tools to extract information from your corpus? What do the results tell you? Why would you want to use a dictionary to annotate your corpus?

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

**Required readings:**


Franzosi, Roberto. TIPS_NLP_Things to do with words Rhetorical analysis Tropes and Figures.pdf

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](https://www.coursera.org/learn/text-mining-analytics/lecture/5RwtX/5-6-how-to-do-sentiment-analysis-with-sentiwordnet)


**Suggested readings:**
You can download SentiWordNet at [http://sentiwordnet.isti.cnr.it/](http://sentiwordnet.isti.cnr.it/)


**Homework 11 (due Sunday March 27, at midnight)**

*Sentiment analysis*

Focusing on the vocabulary and rhetoric of emotions, what does your corpus tell you about emotions? What about Sentiment Analysis? Using the sentiment_analysis_main.py script run the various Sentiment Analysis algorithms. What do the results tell you about the sentiments expressed in your corpus? Which Sentiment Analysis algorithm produces the best results on your corpus?

**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.


Video. Vonnegut, Kurt. [https://www.youtube.com/watch?v=oP3c1h8v2ZQ](https://www.youtube.com/watch?v=oP3c1h8v2ZQ)

Suggested readings:


**Homework 12 (due Sunday April 3, at midnight)**

The shape of stories

According to Kurt Vonnegut stories have “shape” ([https://www.youtube.com/watch?v=oP3c1h8v2ZQ](https://www.youtube.com/watch?v=oP3c1h8v2ZQ)). Does the NLP “shape of stories” tool (shape_of_stories_main.py) applied to your corpus support that claim?

**Part X (Week 13 April 5-7): Dissecting your corpus via the CoNLL table**
Searching the CoNLL table for relationships between words
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
The use of function words, nominalization and passive forms as denial of agency

Software: Stanford CoreNLP, WordNet

Required readings:

Franzosi, Roberto. NLP TIPS files.

Suggested readings:


Homework 13 (due Sunday April 10, at midnight)

Zooming into the CoNLL table

Using the CoNLL table analyzer analyze your corpus in terms of Noun characteristics, Verb modality, Verb tense, Verb voice. What do these terms mean? What did Moretti and Pestre (2015) get out of simple noun and verb statistics? How can you aggregate nouns and verbs using WordNet (WordNet _main.py)? Which WordNet verb categories (top verb synsets) are affected by auxiliaries? What do the results tell you about your corpus? Using the nominalization tool (nominalization_main.py), get a frequency distribution of nominalized verbs. What do the numbers tell you? What do Franzosi et al. (2012) say about nominalization, verb voice, and agency?

Some 300 words are the most frequently words used in the English language. This set of words are often called “junk words” or “stop words”: pronouns, prepositions, articles, conjunctions, and auxiliary verbs. Does your corpus comply to this frequency distribution of words? These words are routinely discarded in computational linguistics analyses. But what do Pennebaker et al. say about pronouns?

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

Week 14: April 12-14

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

Using Gender Guesser for gender attribution: Who wrote this text?

Required readings:
Gender Guesser [http://www.hackerfactor.com/GenderGuesser.php#About](http://www.hackerfactor.com/GenderGuesser.php#About)


**Suggested readings:**


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


**Homework 14 (due Sunday April 17, at midnight)**

*A question of style*

Take a closer look at the writing style of your corpus. Using a variety of tools meant to detect style (easily grouped together under the script `style_analysis_main.py`), analyze your corpus for text readability (at what grade level is your text written?) and sentence complexity. How do character and word n-grams affect style? How do pronouns, nouns, verbs, affect style? Are there gender differences in writing? How does vocabulary affect style? Approach these questions with the range of tools available in the NLP suite and GenderGuesser ([http://www.hackerfactor.com/GenderGuesser.php#About](http://www.hackerfactor.com/GenderGuesser.php#About)).

**Epilogue (Week 15, April 19-21): Digital humanities: A game changer?**

*On visual rhetoric*

**Required readings:**


**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._


**Homework 15 (due Sunday April 24, at midnight)**

**NLP tools and your corpus: The most significant findings**

Time to put it all together. What have you learned about your corpus using NLP tools? Have some consistent themes emerged? Did some tools provide more help than others in bringing out patterns in your data?
In this summary paper, do NOT simply copy and paste results from each homework/presentation. Try to write a coherent story. This may require dropping the results of some NLP tool. After all, you have submitted every homework on every tool; so, there isn’t really a need to submit results from every single tool just to show that you can do it. If you decide not to report the results from a specific tool, you can add a footnote as to why you did that (e.g., because it basically supports the same findings of other tools; or… a specific tool provides slightly different results …; or the tool provides irrelevant and misleading results, e.g., DBpedia and YAGO in annotating folktales). You may also wish to rerun some analyses in light of what you now know.

**Be succinct!!! The paper should be around 2,500 (max!) words in length – about 7 double-space pages max! – excluding visuals.**