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

Emory University

Roberto Franzosi

Office Room No. 212 Tarbutton Hall
Email rfranzo@emory.edu
Office Hours Tu-Th 9:15-11:00 or by appointment

Lectures Tu-Th 8:00-9:15

DEADLINES AND IMPORTANT DATES
First day of class January 26
Last day of class April 29
Homework Due each Sunday at midnight
Readings Due each Thursday
Presentations Due each Thursday, starting on week 5

Table of Contents

COURSE OBJECTIVES ................................................................................................................ 5
Tools of analysis and visualization of text data .............................................................. 5
Learning the language of Natural Language Processing (NLP) ...................................... 5
Big data/small data .............................................................................................................. 5
Visualization and a world of beauty. A game changer? .............................................. 5
Why should YOU take this course: Learning outcomes ........................................ 6
446WR fulfills the writing requirement ........................................................................... 6
Welcome to the 21st century! .............................................................................................. 7
Learning outcomes ........................................................................................................... 7
  Ongoing measurement of learning outcomes .............................................................. 7
Is this a course for YOU? .................................................................................................. 7
  No prerequisites .............................................................................................................. 7
  No prerequisites but... A hard course? ........................................................................ 7
    GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders ...................... 8
    All you need to do is press buttons! ........................................................................... 8
The scariest (and perhaps most frustrating) part is to download and install software .... 8
  All freeware software .................................................................................................. 9
  First things first: What the unzipped NLP Suite looks like ..................................... 9
No Python, No Java, No NLP Suite: Nine Easy steps to install everything! .............................. 10
You also need a corpus .................................................................................................................. 10
What is a corpus? .......................................................................................................................... 10
What types of files do you need for your corpus (csv, pdf, docx)? Txt! ................................. 10
And where would you get this corpus? Option 1: Work on corpora that I provide ............. 11
And where would you get this corpus? Option 2: Get your own corpus .............................. 11
Web scraping .................................................................................................................................. 12
Weekly homework assignments .................................................................................................. 14
Homework rubrics ....................................................................................................................... 14

GRADING ......................................................................................................................................... 14
Participation (5%) ......................................................................................................................... 14
Presentations (10%) – Starting on week 5 .................................................................................. 14
Quizzes (10%) ............................................................................................................................... 14
Homework (45%) .......................................................................................................................... 14
Final paper (30%) .......................................................................................................................... 15
Bonus points .................................................................................................................................... 15

HONOR CODE ............................................................................................................................... 15

WEEKLY HOMEWORK ...................................................................................................................... 15
Homework 1 (due Sunday January 31, at midnight) ................................................................. 16
Homework 2 (due Sunday February 7, at midnight) ................................................................. 16
Homework 3 (due Sunday February 14, at midnight) ............................................................... 16
Homework 4 (due Sunday February 21, at midnight) ............................................................... 16
Homework 5 (due Sunday February 28, at midnight) ............................................................... 16
Homework 6 (due Sunday March 7, at midnight) ..................................................................... 17
Homework 7 (due Sunday March 14, at midnight) .................................................................. 17
Homework 8 (due Sunday March 21, at midnight) .................................................................. 17
Homework 9 (due Sunday March 28, at midnight) .................................................................. 17
Homework 10 (due Sunday April 4, at midnight) .................................................................... 17
Homework 11 (due Sunday April 11, at midnight) ................................................................... 17
Homework 12 (due Sunday April 18, at midnight) ................................................................... 18
Homework 13 (due Sunday April 25, at midnight) ................................................................... 18
Homework 14 (due Sunday May 2, at midnight) ..................................................................... 18

WEEKLY TOPICS & READINGS ...................................................................................................... 18
Required & suggested readings ..................................................................................................... 18
Readings due on Thursdays (Breakout rooms) .......................................................................... 19
Where will you find the readings? ................................................................................................. 19
Introduction (Week 1, January 26-28) ....................................................................................... 19
Big data and “distant reading” ..................................................................................................... 19
Digital humanities: What is it? ..................................................................................................... 19
Preparing your corpus for “distant reading” ............................................................................. 19
Becoming familiar with the suite of Java and Python NLP tools ............................................. 19
Part 1 (Week 2, February 2-4): Visualizing Words .................................................................. 21
Word clouds .................................................................................................................................... 21
Visualization in Digital humanities .............................................................................................. 21
HTML annotated files: Wikipedia, DBpedia, Yago, dictionaries ............................................. 22
Excel charts (with hover-over effects) ......................................................................................... 22
Network graphs: Mapping relations.......................................................... 23
Part II (Week 3, February 9-11): What’s in your corpus? A sweeping look .......... 23
  What are the topics? Topic modeling .................................................. 23
  Is there dialogue? .............................................................................. 23
  Are there people and organizations? .................................................... 23
  Are there locations? ........................................................................... 23
  Are there times? .................................................................................. 23
  Does nature appear? ........................................................................... 23
  Do verbs and nouns cluster in specific classes? ..................................... 24
Part III (Week 4, February 16-18): NLP (Natural Language Processing): Basic language .... 25
  Sentence splitter, tokenizer, lemmatizer, parser .................................... 25
  Stanford CoreNLP and the CoNLL table ............................................... 25
Part IV (Weeks 5-6, February 23-25-March 2-4): Searching your corpus ............ 27
  A closer look at the CoNLL table ......................................................... 27
  Meet the NER, POSTAG, DEPREL .................................................... 27
  Searches by words’ grammatical roles (POSTAG) ............................... 27
  Searches by syntactic relations (DEPREL) .......................................... 27
  Searches by NER values (Named Entity Recognition) ........................... 27
  Google N-grams Viewer and Culturomics ......................................... 28
  N-grams searches in the NLP Suite .................................................... 28
  Word co-occurrences searches ............................................................ 28
  Single words/collocations searches ..................................................... 28
Part V (Week 7, March 9-11): From text to maps ......................................... 29
  Using CoNLL NER information to map locations ................................. 29
  Geocoding ......................................................................................... 29
  Visualizing time and space .................................................................. 29
Part VI (Weeks 8-9, March 16-18, 23-25): Narrative .................................... 30
  The 5 Ws of Narrative: Who does What, When, Where, and Why ......... 31
  Of narrative time: story and plot .......................................................... 32
Part VII (Weeks 10-11, March 30-April 1, April 6-8): The world of emotions .... 32
  The words of emotions ....................................................................... 32
  The rhetoric of emotions: punctuation and repetition ......................... 33
  Sentiment Analysis: Capturing the feelings conveyed in the writing ....... 33
  The “shape” of stories ........................................................................ 34
Part VIII (Week 12 April 13-15: Dissecting your corpus via the CoNLL table ...... 35
  Noun density and noun types ............................................................. 35
  Verb modality: Ability, possibility, permission, and obligation ................ 35
  Verb tense: past, future, gerundive ....................................................... 35
  Verb voice: Active and passive verb forms ......................................... 35
  Nominalization .................................................................................... 35
  Function words (“junk” words or “stop” words): pronouns, prepositions, articles,
  conjunctions, and auxiliary verbs ...................................................... 35
  Pronouns and Coreference resolution ................................................. 35
Part IX (Week 13 April 20-22): A question of style ....................................... 36
  Back to the CoNLL table and what it reveals about style ...................... 36
  Text readability: What grade level does a text require to be comprehensible? .... 36
Sentence complexity: Measuring and visualizing linguistic complexity .................. 36
Analyzing vocabulary .................................................................................................. 36
N-grams ...................................................................................................................... 36
Using Gender Guesser for gender attribution: Who wrote this text?......................... 36
Part X (Week 14 April 27-29): Digital humanities: A game changer? ....................... 37
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…
… 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 (not to worry! This includes graphical displays). But the rest is easy.
GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders

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

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

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

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

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

The scariest (and perhaps most frustrating) part is to download and install software

But you know what? That will all happen in week 1 and with a great deal of help to get you going.
All freeware software

We will only use freeware software.

First things first: What the unzipped NLP Suite looks like

There is nothing to be installed. You just need to download and unzip the NLP Suite.zip file. And once unzipped, this is what the NLP Suite directory will look like, with each subdirectory containing required files. Let’s review them briefly.

Name

- config
- lib
- reminders
- sampleData
- sic
- TIPS

Read very carefully the readme files!!!

config. After you start using the Suite, the config subdirectory will contain a set of txt files where the default Input/Output options of each script will have been saved. This will avoid having to reenter input and output filenames/directories every time you run a script.

lib. The lib subdirectory contains a handful of library files required by some scripts.

reminders. The subdirectory contains a csv file that lists all the reminders available for the various scripts and with their status (On or Off, as set by the user).
**sampleData.** The subdirectory contains a handful of txt files for the user to test scripts.

**src.** The subdirectory contains over 120 Java (jar) and Python files that make up the NLP Suite.

**TIPS.** Like the py files, the number of TIPS files is constantly expanding, currently numbering 100. The TIPS files provide detailed online help on the various NLP tools.

```
<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python TIPS...</td>
<td>02/02/2023</td>
<td>1.5MB</td>
</tr>
<tr>
<td>Java TIPS...</td>
<td>02/02/2023</td>
<td>2.0MB</td>
</tr>
<tr>
<td>NLP_suite.zip</td>
<td>02/02/2023</td>
<td>10MB</td>
</tr>
</tbody>
</table>
```

**No Python, No Java, No NLP Suite: Nine Easy steps to install everything!**

This is where it gets a bit more complicated. To the very least you need to install Python and Java, the computer languages the NLP Suite is written in. No Java and, especially, no Python, no NLP Suite.

**But… here is the good news. Read the NLP SUITE Readme 4 Windows.pdf or the NLP SUITE Readme 4 MAC.pdf. These Readme files outline 9 easy steps with screenshots and explanations. Even I could do it, following the steps!!!**

The Readme files are in the NLP_test.zip and on CANVAS under files.

**You also need a corpus**

**What is a corpus?**

You will need to come to the course in week 1 with a larg(ish) set of TEXT documents (that’s what a corpus is), the more the better (at least 20 or 30 documents, such as 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.
And where would you get this corpus? Option 1: Work on corpora that I provide

I have several text corpora that I have collected and that you can analyze. You can analyze each corpus individually or in a team of max 3 students. If you work in teams, you are expected to submit only one joint homework assignment a week but… you are, obviously, expected to do more extensive work, whether of scholarly research into the topic of a specific corpus or computational work. But, individually or in team, the goal is the same: arrive at the end of term in your final paper with publishable quality work for journal submission. Depending upon the quality of your work, you will co-authors in joint journal submissions.

   These are 700 personal stories by African males and females affected by HIV. The project would be together with a professor from the School of Public Health.

2. **Health stories from web blogs**
   We have collected over 200 personal narratives of health taken from various web blogs that we have used to analyze the shape of stories.

3. **Interviews on social mobility in the UK: Goldthorpe vs. Savage. Can NLP settle the controversy?**
   John Goldthorpe and Mike Savage, two very famous English sociologists, quarreled over social mobility and analyzed differently some 200 in-depth interviews using hand-coding of these interviews. Can we settle the quarrel with NLP tools?


5. **Analyzing folktales from China, England, Germany, India, Middle East, and Russia**
   Are there national differences in folktales? Is the sense of magic and wonder only part of Eastern cultures? Is the accusation against the German Grimm brothers fair that their stories were NOT for children’s ears, with too much sex and violence?

   Which criteria do reviewers use in evaluating best-selling books? Does an author’s race/ethnic background lay any role in this? How are book reviews constructed as a literary genre? Phillipa Chong addresses these questions using a qualitative approach to reviews. Can NLP help us address those same questions? We have scraped some 100 NYT book reviews with which to try to answer these questions. We would need to scrape more reviews but… I have already written the paper!

And where would you get this corpus? Option 2: Get your own corpus

1. **blogs**
2. newspaper articles
3. US Congress bills (https://www.congress.gov/)
5. corporate/university mission statements
6. 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)
7. oral history archives; see the list provided by the Oral History Association, (http://www.oralhistory.org/centers-and-collections/)
8. transcribed in-depth interviews
10. NYT book reviews; see the NYT API (http://developer.nytimes.com/docs/books_api/)
11. song lyrics; see, for example, the collection provided by AZlyrics (http://www.azlyrics.com/a/archive.html)
12. 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.oocl.org/oaister.en.html)
13. diaries & autobiographies
14. 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. 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.
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. These texts are mostly chosen by the individual students, except for some assigned texts (e.g., the short story “Professor Dumbass”). 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, 2 to 5 pages of writing are expected every week. But the amount of writing is expected to 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 weekly on CANVAS.**

**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 (10%) – Starting on week 5.** Students will be asked to make in-class presentations of their work. **Presentations will start on week 5. 10-15 minutes max in Power Point!!!**

**Quizzes (10%)** – There will 2 unannounced multiple-choice and short answer questions quizzes covering reading material and lectures.

**Homework (45%).** 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;**
b. **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.

Homework will be graded following the rubric and returned within a week of due date.

**Final paper (30%).** You will be expected to write a final research paper based on the analysis of corpus data of your choice. You are welcome to organize your paper in the standard format – Introduction, Literature Review, Data & Methods, Empirical Results, Conclusions, Bibliography – but you are also encouraged to experiment with creative writing (as long as all relevant information of the standard format is still provided). Much of the analysis will have been carried out for the weekly assignments. But try not to simply copy and paste results from each homework. Try to write a coherent story. You may decide not to include the results of every single NLP tool you learned. After all, you have submitted every homework on every tool; so there isn’t really a necessity to submit results from every single tool just to show that you can do it. If you really want to, you can add a footnote as to why you did not report the results from a specific tool (e.g., because it basically supports the same findings of other tools; or… a specific tool provides slightly different results …).

You should aim to write a publishable quality paper. After all, the paper must live up to the expectations of the W of the course, fulfilling the Writing requirement.

The paper should include plots, charts, graphs, and links to dynamic visualizations.

The paper should be around 5 thousand words in length excluding visuals.

**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.** Programming work can also be carried out instead of some weekly homework.

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://www.sph.emory.edu/cms/current_students/enrollment_services/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.

Homework 1 (due Sunday January 31, at midnight)

1. Provide screenshots of successful installation of software on your computer.
2. 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. Separately, write two pages on “distant reading” and the “digital humanities.” What do these concepts mean? Why distant? What are the pros and cons of distant reading?

Homework 2 (due Sunday February 7, at midnight)

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

Homework 3 (due Sunday February 14, at midnight)

Using the topic modeling tools Gensim and Mallet, analyze your corpus for significant topics in your corpus. Do Gensim and Mallet correctly categorize your corpus? Which tool performs better? What else is in your corpus? Dialogue, people, locations, times, noun and verb classes?

Homework 4 (due Sunday February 21, at midnight)

1. 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. From the CoNLL table, compute the Sentence table. 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.

Homework 5 (due Sunday February 28, at midnight)

Write a five-page report on the results of using POSTAG, DEPREL, NER SEARCH TOOLS on your corpus (CoNLL_search_main.py), and 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?
Homework 6 (due Sunday March 7, at midnight)

Write a five-page report on the results of the SEARCH TOOLS encountered this week and applied to your corpus (NGrams_CoOccurrences_main.py). 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?

Homework 7 (due Sunday March 14, at midnight)

Using the NER information of your corpus (if your corpus contains locations) and Professor Dumbass story in the sampleData subfolder of NLP, 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)?

Homework 8 (due Sunday March 21, at midnight)

March 15-19 “no assignment week” NO HOMEWORK DUE!

Homework 9 (due Sunday March 28, at midnight)

Using the SVO extractor tool (SVO_main.py), analyze your corpus to extract the Who, the What, the When and Where of narrative. Run the tool with and without coreference resolution (via Stanford CoreNLP). What difference does it make? Finally, use the resolution GUI to resolve the 35% cases not dealt with automatically. Again… what difference does it make?

Using the Story & Plot GUI based on Stanford CoreNLP SUTIME and NER annotators (Person and Organization), can you map the different “scenes” of a story, with characters acting in different times and locations?

Homework 10 (due Sunday April 4, at midnight)

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?

Homework 11 (due Sunday April 11, at midnight)

According to Kurt Vonnegut stories have “shape” (https://www.youtube.com/watch?v=oP3e1h8v2ZQ). Does the NLP “shape of stories” tool (shape_of_stories_main.py) applied to your corpus support that claim?
**Homework 12 (due Sunday April 18, at midnight)**

Using the CoNLL_table_analyzer_main.py tool, 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? Use the nominalization tool (nominalization_main.py) to 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? How can you aggregate nouns and verbs using WordNet (WordNet_main.py)? What do the results tell you about your corpus?

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. Using the CoNLL_function_words_main.py, 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?

What role do pronouns play in coreference resolution? Using the coReference_resolution_main.py script, please analyze your corpus. Do personal pronouns get “resolved” by the Stanford CoreNLP algorithm? What does it mean to “resolve” coreferences? If you analyze your unresolved and resolved corpus data, is there a difference in the result? Do results improve? How much extra work did you have to do to completely resolve pronominal coreferences manually?

**Homework 13 (due Sunday April 25, at midnight)**

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).

**Homework 14 (due Sunday May 2, at midnight)**

No homework due

**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).

Readings due on Thursdays (Breakout rooms)

Readings are expected to have been carried out by each Thursday. They will be discussed in small groups in Breakout rooms who will then report back to the class as a whole. Two important questions will be discussed:

1. What do the readings say?
2. Why are the week’s NLP tools important? What do they allow you to say about a (your) corpus?

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 26-28)

Big data and “distant reading”
Digital humanities: What is it?
NLP: What is it?
Preparing your corpus for “distant reading”
Becoming familiar with the suite of Java and Python NLP tools

What does your corpus look like?

Required readings:
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)

**Part I (Week 2, February 2-4): Visualizing Words**

**Word clouds**

*Visualization in Digital humanities*  
(From Jänicke et al. 2015)

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

Required readings:


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:**


**HTML annotated files:** Wikipedia, DBpedia, Yago, dictionaries

---

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

**Required readings:**

Franzosi, Roberto. NLP TIPS files.

https://www.dbpedia-spotlight.org/demo/
https://gate.d5.mpi-inf.mpg.de/webyago3spotlxComp/SvgBrowser/

**Excel charts (with hover-over effects)**
Network graphs: Mapping relations

Required readings:
Franzosi, Roberto. NLP TIPS files.

Part II (Week 3, February 9-11): What’s in your corpus? A sweeping look

What are the topics? Topic modeling

via Gensim and Mallet

Is there dialogue?

Use CoreNLP quote annotator

Are there people and organizations?

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

Are there locations?

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

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

Are there times?

Use CoreNLP NER normalized time annotator to extract standardized temporal expressions

Does nature appear?

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

Use DBpedia/YAGO to annotate texts
Do verbs and nouns cluster in specific classes?

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

Software: Mallet & Gensim

Required readings:

Puccini, Giacomo. 1900. *Tosca*, “E lucevan le stelle” Luciano Pavarotti performing in Teatro dell' Opera in Rome, 2000; Plácido Domingo conducting. [https://www.youtube.com/watch?v=TU5roitYI1s](https://www.youtube.com/watch?v=TU5roitYI1s)

**E lucevan le stelle (English translation)**

And the stars were shining,
And the earth was scented.
The gate of the garden creaked
And a footstep grazed the sand...
Fragrant, she entered
And fell into my arms.

Oh, sweet kisses and languorous caresses,
While feverishly I stripped the beautiful form of its veils!
Forever, my dream of love has vanished.
That moment has fled, and I die in desperation.
And I die in desperation!
And I never before loved life so much,
Loved life so much!

Both Bunin and Faulkner available in txt format in sampleData under NLP.

Franzosi, Roberto. NLP TIPS files.


For an interesting paper 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](https://www.youtube.com/watch?v=WSbgixdC9g8)

*Part III (Week 4, February 16-18): NLP (Natural Language Processing): Basic language*

*Sentence splitter, tokenizer, lemmatizer, parser*

*Stanford CoreNLP and the CoNLL table*
Software: Stanford CoreNLP

Required readings:

Top 20 free software for Text Analysis, Text Mining, Text Analytics
http://www.predictiv analyticstoday.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.


**Part IV (Weeks 5-6, February 23-25-March 2-4): Searching your corpus**

**Week 5: February 23-25**

*A closer look at the CoNLL table*

*Meet the NER, POSTAG, DEPREL*

*Searches by words’ grammatical roles (POSTAG)*

*Searches by syntactic relations (DEPREL)*

*Searches by NER values (Named Entity Recognition)*

**Software: Stanford CoreNLP**

**Required readings:**

Franzosi, Roberto. NLP TIPS files.

**Week 6: March 2-4**
**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**

![Graph showing N-gram frequency over time](image)

**Required readings:**

Franzosi, Roberto. NLP TIPS files.

**Become familiar with the basic language of culturomics!**


**Suggested readings:**


**Part V (Week 7, March 9-11): 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

*Part VI (Weeks 8-9, March 16-18, 23-25): Narrative*

*Week 8: March 16-18*
The 5 Ws of Narrative: Who does What, When, Where, and Why

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:


Suggested readings:


**Week 9: March 23-25**

*Of narrative time: story and plot*

**Required readings:**

Franzosi, Roberto. NLP TIPS files.


**Suggested readings:**


**Part VII (Weeks 10-11, March 30-April 1, April 6-8): The world of emotions**

**Week 10: March 30**

*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.

*Required readings:*


**Week 10: April 1**

*Sentiment Analysis: Capturing the feelings conveyed in the writing*

*Required readings:*

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/)
Week 11: April 6-8

The “shape” of stories

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:


Part VIII (Week 12 April 13-15: Dissecting your corpus via the CoNLL table

![Image of CoNLL table]

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

Software: Stanford CoreNLP, WordNet

Required readings:

Franzosi, Roberto. NLP TIPS files.


Suggested readings:


Part IX (Week 13 April 20–22): A question of style

Back to the CoNLL table and what it reveals about style
The use of function words, nominalization and passive forms as denial of agency

Text readability: What grade level does a text require to be comprehensible?
Sentence complexity: Measuring and visualizing linguistic complexity
Analyzing vocabulary
N-grams
Using Gender Guesser for gender attribution: Who wrote this text?

Required readings:

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

Suggested readings:


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


For visual rhetoric

Required readings:

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.


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.


