“[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

Robert Franzosi

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

Lectures
Tu-Th 4:00-5:15 Callaway Center S101

DEADLINES AND IMPORTANT DATES

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

Table of Contents

COURSE OBJECTIVES .............................................................................................................................................. 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
Welcome to the 21st century! .............................................................................................................................. 6
Learning outcomes .................................................................................................................................................. 6
Ongoing measurement of learning outcomes ..................................................................................................... 7
Is this a course for YOU? .................................................................................................................................... 7
No prerequisites ..................................................................................................................................................... 7
GUI (Graphical User Interface): HELP, Read Me, TIPS, Reminders ................................................................. 7
All you need to do is press buttons! ................................................................................................................... 7
Download and install the NLP Suite and other software ........................................... 8
   NLP Suite welcome GUI ......................................................................................... 8
You need a work partner .............................................................................................. 9
   You need a corpus ......................................................................................................... 9
Homework ...................................................................................................................... 10
Weekly presentations ..................................................................................................... 10
   Presentation rubrics ..................................................................................................... 10
GRADING ....................................................................................................................... 10
   Participation (25%) ..................................................................................................... 10
   Presentations (75%) – Starting on week 3 ................................................................. 10
   Bonus points ............................................................................................................... 11
HONOR CODE ............................................................................................................... 11
WEEKLY TOPICS & READINGS ................................................................................. 11
   Required & suggested readings .................................................................................... 11
   Where will you find the readings? ................................................................................ 11
Introduction (Week 1, January 11-13) ........................................................................ 11
   Big data and “distant reading” .................................................................................... 11
   Digital humanities: What is it? .................................................................................... 11
   Becoming familiar with the suite of Java and Python NLP tools .............................. 11
Part I (Week 2, January 18-20): Corpus Statistics and Words Visualization .............. 12
   Corpus Statistics ........................................................................................................ 12
   Visualization in Digital humanities ............................................................................ 13
   Word clouds ................................................................................................................ 13
   Excel charts (with hover-over effects) ........................................................................ 13
   Network graphs: Mapping relations ......................................................................... 14
   Knowledge graphs (KG) and HTML annotated files .............................................. 14
   Maps: Space (and time) ............................................................................................. 15
Part II (Week 3, January 25-27): Topic Modeling & Word2Vec .................................. 16
   What are the topics in your corpus? .......................................................................... 16
   Topic modeling via Gensim and Mallet .................................................................. 16
   Word2Vec ................................................................................................................... 16
Part III (Week 4, February 1-3): NLP (Natural Language Processing): Basic language .. 17
   Sentence splitter, tokenizer, lemmatizer, parser ..................................................... 18
   The Stanford CoreNLP parsers ................................................................................. 18
   Meet the CoNLL table ............................................................................................... 18
Part IV (Week 5, February 8-10): Named Entity Recognition (NER) and CoreNLP annotators ................................................................. 19
   A closer look at the CoNLL table: Meet the NER, POSTAG, DEPREL tags .......... 19
   Stanford CoreNLP annotators ................................................................................. 19
   Is there dialogue? ....................................................................................................... 19
   Are there people and organizations and differences in gender distribution? .......... 19
   Are there geographical locations? ............................................................................ 19
   Are there times? ......................................................................................................... 19
   Using WordNet: Does nature appear? ..................................................................... 20
   Using WordNet: Do nouns and verbs cluster in specific classes? ......................... 20
Part V (Week 6, February 15-17): From text to maps .................................................. 20
Using CoNLL NER information to map locations ............................................. 20
Geocoding ........................................................................................................ 20
Visualizing time and space ............................................................................. 20
Part VI (Week 7, February 22-24): Narrative and the 5 Ws ..................................... 21
SVO Extraction & Visualization ..................................................................... 22
Stanford CoreNLP enhanced dependencies parser ......................................... 22
SENNA .............................................................................................................. 22
Stanford CoreNLP OpenIE .............................................................................. 22
Part VII (Weeks 8-9, March 1-3, March 8-10): N-grams, co-occurrences, culturomics ... 23
A closer look at N-grams ................................................................................. 24
Google N-grams Viewer and Culturomics ....................................................... 24
N-grams searches in the NLP Suite ................................................................. 24
Word co-occurrences searches ....................................................................... 24
Single words/collocations searches ............................................................... 24
Part VIII (Week 10, March 15-17): Knowledge-base systems (DBpedia and YAGO) .... 25
DBpedia ............................................................................................................. 25
YAGO ................................................................................................................. 25
dictionary-based annotation ......................................................................... 25
html files .......................................................................................................... 25
Part IX (Weeks 11-12, March 22-24, March 29-31): The world of emotions .......... 26
Sentiment Analysis: Capturing the feelings conveyed in the writing ............... 26
WordNet ............................................................................................................ 26
YAGO ................................................................................................................. 26
ANEW ................................................................................................................. 26
Hedonometer .................................................................................................... 26
SentiWordNet .................................................................................................. 26
Stanford CoreNLP sentiment analysis annotator ............................................ 26
VADER ................................................................................................................. 26
The “shape” of stories ..................................................................................... 27
Data reduction algorithms: Hierarchical Clustering (HC), Singular Value Decomposition (SVD), Non-Negative Matrix Factorization (NMF) ...................................................... 27
Part X (Week 13 April 5-7): Dissecting your corpus via the CoNLL table ............... 28
Noun density and noun types ....................................................................... 28
Verb modality: Ability, possibility, permission, and obligation ....................... 28
Verb tense: past, future, gerundive ............................................................... 28
Verb voice: Active and passive verb forms ................................................... 28
Function words (“junk” words or “stop” words): pronouns, prepositions, articles, conjunctions, and auxiliary verbs ...................................................... 28
Pronouns and Coreference resolution ............................................................ 28
Part XI (Weeks 14-15, April 12-14, April 19-21): A question of style .................... 29
Back to the CoNLL table and what it reveals about style ................................ 29
Text readability: What grade level does a text require to be comprehensible? .... 29
Sentence complexity: Measuring and visualizing linguistic complexity .......... 29
Analyzing vocabulary .................................................................................... 29
N-grams and style .......................................................................................... 29
The use of function words, nominalization and passive forms as denial of agency ... 29
Using Gender Guesser for gender attribution: Who wrote this text? ........................................... 29
Epilogue (Week 16, December 7): Digital humanities: A game changer? ............................. 31
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**

*Welcome to the 21st century!*

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

*Learning outcomes*

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

1. Understand the concepts of big data, Natural Language Processing (NLP), Artificial
Intelligence, machine learning…
2. Use a variety of NLP tools and what they can do
3. Use a variety of data visualization tools, drawing geographic maps, network graphs, charts …
4. Make public presentations before an audience
5. Write research reports

Ongoing measurement of learning outcomes

Learning outcomes will be assessed every week through weekly presentations.

IS THIS A COURSE FOR YOU?

No prerequisites

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

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

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

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

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

HELP, Read Me, Videos, TIPS, Reminders buttons are all at your fingertips. Hard to screw up!
Example of TIPS file… TIPS files, at least the longer ones, even come with a Table of Contents.

```
Stanford CoreNLP Parser

What is a parser? ................................................................. 1
Free open-source parsers .................................................. 1
The Stanford CoreNLP parsers ........................................... 2
System requirements ....................................................... 2
Java .................................................................................. 2
Input: The CoNLL table ...................................................... 2
Output: The neural-network dependency parser and named tags .................................................. 3
Faulty results? .................................................................. 3
```

**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 3.

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

You need a corpus

We have several text corpora that you can analyze.

**Gay men project**

- 376 personal narratives from gay men from 37 different countries
- **The Harry Potter books**
  J. K. Rowling’s collection of 7 Harry Potter books
- **US presidential speeches** ([https://www.presidency.uesb.edu/](https://www.presidency.uesb.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)
- **New York Times best-selling book reviews**
  Some 1300 NYT best-selling book reviews
- **Folktales**
A collection of hundreds of cross-national English, German, Chinese, Arabic, and Indian folktales

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

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

Weekly presentations

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

Presentation rubrics

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

GRADING

Grading will be based on the following items:

Participation (25%). You are expected to attend classes regularly (attendance is enforced through a sign-up sheet) and contribute to discussion.
Presentations (75%) – Starting on week 3 student teams will make in-class presentation of their work. 10-15 minutes max in Power Point with the use of graphical displays.
Presentations will cover an overview of the corpus (what is the corpus about? number of documents, of sentences per document, linguistic domain as shown by the distribution of words) and the most significant results using the tools learned by the time of the presentation (from n-grams, to topic modeling, CoreNLP annotators – gender, normalized dates, quote – knowledge-base and dictionary annotators, SVO extractors, sentiment, style, and more… What are the pros and cons, strengths and limits of the NLP tools used? As the semester progresses and students learn more NLP tools, repeated teams’ presentations are expected to provide both broader and more in-depth analyses of the corpora.

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

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

HONOR CODE

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

WEEKLY TOPICS & READINGS

Required & suggested readings

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

Where will you find the readings?

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

TIPS_NLP_Things to do with words NLP approach.pdf


Suggested readings:


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

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

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

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

File types doc, docx, rtf, txt, pdf) and what to do about it

Corpus Statistics
Get basic statistics about your corpus: number of documents, number of sentences, number of words; Ngrams
Visualization in Digital humanities

Word clouds

See

Wing

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

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:


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

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

Software: Mallet & Gensim
Franzosi, Roberto. NLP TIPS files.


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 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.predictivestatsticstoday.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=mmWRNRb0W0

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

*A closer look at the CoNLL table: Meet the NER, POSTAG, DEPREL tags*

*Stanford CoreNLP annotators*

*Is there dialogue?*

*Are there people and organizations and differences in gender distribution?*

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

*Are there geographical locations?*

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

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

*Are there times?*

Use CoreNLP NER normalized time annotator to extract standardized temporal expressions
Software: Stanford CoreNLP

Using WordNet: Does nature appear?

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

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

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

Required readings:

Franzosi, Roberto. NLP TIPS files.

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

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

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

Required readings:
Franzosi, Roberto. Geocoding TIPS files.


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

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:

Franzosi, Roberto. NLP TIPS files.


Suggested readings:


(Found under Murchú or zip will not zip)


http://verbs.colorado.edu/~mpalmer/dossier/HindiIntro.pdf


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

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

Software: Stanford CoreNLP, Google Ngram Viewer

Required readings:

Franzosi, Roberto. NLP TIPS files.

Become familiar with the basic language of culturomics!

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

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

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

Suggested readings:


Week 9: March 1-3

SPRING BREAK March 8-10 no classes

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

DBpedia
YAGO
Dictionary-based annotation
html files

Required readings:

Franzosi, Roberto. NLP TIPS files.


Suggested readings:

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

Week 11: March 22-24

*The words of emotions*

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

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

*The rhetoric of emotions: punctuation and repetition*

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

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

*WordNet*
*YAGO*
*ANEW*
*Hedonometer*
*SentiWordNet*
*Stanford CoreNLP sentiment analysis annotator*
*VADER*

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

Franzosi, Roberto. NLP TIPS files.

*Suggested readings:*

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


**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:**


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

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

**Software:** Stanford CoreNLP, WordNet

**Required readings:**

Franzosi, Roberto. NLP TIPS files.


Suggested readings:


For an excellent socio-linguistic use of Pennebaker’s work on function words, see:


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

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

Required readings:

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


Suggested readings:


For a state-of-the-art review of authorship attribution, see
Epilogue (Week 16, December 7): Digital humanities: A game changer?

On 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.


Tufté 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.

