# Course Syllabus: Introduction to Natural Language Processing – Revised Jan 23

CSCI E-89B*.*

Harvard Extension Spring 2021

Dates: Jan 25, 2021 – May 15, 2021

Time: Thursday 8:10-10:10pm

Building: NA, Remote only.

Instructor: Ted Kwartler, MBA

Email:

[edwardkwartler@fas.harvard.edu](mailto:edwardkwartler@fas.harvard.edu)

Office Hrs: Available upon request

Optional Lab: TBD

## Important URLs:

**Canvas** (homework submissions and grading)  
[Canvas](https://canvas.harvard.edu/courses/81758)

**Piazza** (class forum for discussions and asking questions)

[Piazza](https://piazza.com/class/kjud8fkxe1u3il)

This term we will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates and your teaching staff. Rather than emailing questions to the teaching staff directly, I encourage you to post your questions on Piazza to benefit everyone.

Keep in mind, Piazza is used to realize distance learning, but it is never intended to be a customer service center, social networking website or the channel for sharing evaluations of teammates, staff or topics (save that for your course evaluations). Further, teaching staff responses are not expected within any given timeframe. In fact, your teaching staff may not respond at all for some topics so that students have to help each other, research and explore on their own and ultimately learn rather than be explicitly told. While some students find this problematic the course goal is to facilitate self-learning and improve confidence as much as direct instruction.

The **Github** repository allows you to get all scripts, PowerPoints and data sets throughout the semester. For those not familiar with github, think of it like a shared drive similar to SharePoint or Dropbox but with added functionality for data and computer science.

*There are no spaces in the link below, they are underscores (“\_”)*

<https://github.com/kwartler/Harvard_NLP_Student>

## Streaming & Video Information:

Lectures will be streamed via zoom, with link in the canvas site.

Students will be able to access the recordings to watch on demand. Recordings are usually posted within 24-48hours after the lecture.

**All lecture video links will be available from the Course Canvas site.**

## Prerequisites:

* Textbook: Text Mining in Practice with R

ISBN-10: 1119282012

* Software: R & R-Studio
  1. This course expects basic understanding of R and requires a non-graded script for assessment.
  2. If you require a refresher for R programming please take a short introduction to R course at Lynda.com, DataQuest.com or DataCamp.com.
* Access to git software to download data sets and class material or ability to download directly from the Internet
* A webcam or other method to record case presentations & upload to the University’s approved site as well as interact with live class sessions if that is your mode of attendance. If you prefer you can share videos directly with teaching staff for privacy.
* Be prepared to obtain a free zoom account to record case presentations
* To avoid disruption please install R and R studio on your local laptop. This requires you to have administration privileges. Further one of the R packages `qdap` requires a java installation which may be challenging on Mac OS. As a backup you may use rstudio.cloud with the Harvard University workspace.

## Course Learning Objectives:

Natural Language Processing (NLP) and Text mining is the art and science of extracting insights from large amounts of natural language.  The course topics will help students add natural language processing techniques to their research, business and data science toolset.  As a technical course with some machine learning elements, limited exposure to programming, graduate level statistics and mathematical theory is needed but the vast majority of the course content will be focused on applying popular text mining methods.  As a result, the target audience may also include qualitative researchers looking to add quantitative analysis to interviews, media and other language based field research as long as participants have some basic R background.

Students will be able to think systematically about how information can be obtained from diverse natural language.   
Students will learn how to implement a variety of popular text mining algorithms in R (a free and open-source software) to identify insights, extract information and measure emotional content.  
Students will learn practical applications of NLP affecting consumers, researchers, and businesses.

## Attendance:

Regular attendance (expressed by watching videos live or asynchronously) and remote participation (expressed by interacting on the class forum) is essential to the successful completion of this course. If applicable, attendance will be taken regularly for on campus sessions and forum participation will be monitored for remote participants. Remote participation is measured by questions asked, posts, post views and responses to posts. You are responsible for material covered in class even if you have not attended class or watched the recorded lectures. For classes on campus, given the amount of information covered, missing more than 1 class session for any reason may result in an automatic reduction in course grade. Unsatisfactory attendance may result in a failing grade. For remote asynchronous participants, skipping videos and not participating on forums may impact both your assignments sophistication and also your participation grade. In either class delivery mode, you should plan on spending at least three hours of independent study for each hour of class attendance.

## Code of conduct:

This course expects you to uphold and report violations of the Extension School code of conduct found [here](https://www.extension.harvard.edu/resources-policies/student-conduct). Further, all assignments are the responsibility of each *individual* pupil unless assigned as a group assignment. Utilizing the class forum, online resources, teaching assistants, and the class professor to ask questions is (of course) acceptable but copying another peer’s work is considered a violation of the University code of conduct.

You are responsible for understanding Harvard Extension School policies on academic integrity ([www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity](http://www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity)) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting "the wrong draft", or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism ([www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism](http://www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism)), where you'll find links to the Harvard Guide to Using Sources and two, free, online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.  
  
Accessibility  
The Extension School is committed to providing an accessible academic community. The Disability Services Office offers a variety of accommodations and services to students with documented disabilities. Please visit [www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility](http://www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility) for more information.

## Grading:

A course grade will be assigned on the basis of student performance on case studies, a written assignment, and attendance and participation.

Case material is accepted up to 12 hours late. Any work submitted after the deadline but before 12 additional hours will be penalized 1 letter grade. After 12 hours no late submissions will be accepted under ANY circumstances. Pupils are expected to manage their own time and submit their work accordingly. Failure to submit submissions through the University approved portal by the assignment deadline will be considered late and not accepted. Submissions to any other location will not be accepted.

* Class participation, and online forum participation 15% of final grade. The course is a collaborative learning environment. The expectation is that *all* students will view, comment, post interesting/current data science and business topics or news and ask questions. *Class participation is not free credit. If students do not contribute, they will not receive class participation credit.* Remote, asynchronous students are expected to contribute to the forum, by asking questions, posting relevant articles or answering others questions.

**Graduate Student Grading**

* Class Participation 15%
* Case I 25% of final grade  **– Fan Engagement NBA Tweets**
* Case II 25% of final grade **– Investigative Journalism Research – Budget Analysis**
* Case III 25% of the final grade **– Political Discourse**
* Written assignment 10% of final grade

**Undergraduate Student Grading**

* Class Participation 25%
* Case I 25% **– Fan Engagement NBA Tweets**
* Case II 25% **– Investigative Journalism Research**
* Written assignment 25% of final grade

## Writing Assignment

A portion of the final grade will be determined by the quality and completeness of a 900 to 1200 word ***essay concerning a personal code of conduct for using natural language processing ethically either in research or business***. For professionals in the class, this may mean articulating a justification for moral business applications using this technology, identifying aspects of the technology one is not comfortable with and identifying possible objections to demonstrate robust thoughtfulness. For researchers and students, this essay may demonstrate introspection of how society, is shaped by this technology and its possible missteps or even ways to perform research in an ethical manner utilizing NLP methods.

Example questions to spur creative reflection include (but are not limited to):

* Does publishing a research paper using public data contribute to amplifying messages in the case of racism, sexism etc? If so, how would you navigate conducting research in these areas or would you deviate from these topics?
* Is placing smart speakers employing natural language processing in a home ethical in all circumstances? Would you as a professional want to work on a project to create this technology in other areas such as in the workplace? Would an employer have the moral justification to spy on their employees if a smart speaker is present?

While defining an ethical framework can be a personal matter, the organization and robustness of your argument along with supporting statements to the argument are subject to evaluation. It is not the case that all ethical actions are relative or that ethical considerations are incapable of objective evaluation. Further the level of sophistication you demonstrate in understanding the issue discussed, addressing applicable opposing viewpoints, actions stakeholders can take to mitigate issues and the logical structure of your essay will impact your grade. Lastly, primary source philosophical paradigms, not mere opinions should be used as a foundation for your logical construction of what is ethical in a data mining and business context.

Each page should have a header with a clear label including the author, date, page number and title. As a personal reflection paper concerning ethics, APA or similar citation method is *not* necessary.

## Case Presentations

Each student will work on 3 case studies individually (or 2 as an undergraduate). Cases will involve using text to apply various methods and draw out insights and conclusions. Each case will have the following work artifacts:

1. Case I: Fan engagement in the NBA
   1. 500-1000 word “abstract” or summary as if the research is to be included in an academic journal. Describe the text used, processing steps taken, insights identified.
   2. R script, markdown or notebook supporting for importing, processing, applying methods and creating any visuals described in the 500-1000 word document.
2. Case II: Investigative Journalism – Budget Analysis
   1. 500-1000 word “abstract” or summary as if the research is to be included in an academic journal. Describe the text used, processing steps taken, insights identified.
   2. R script, markdown or notebook supporting for importing, processing, applying methods and creating any visuals described in the 500-1000 word document.
3. Case III: Political Discourse, Document Classification
   1. Maximum 10min voice recorded slide presentation uploaded to youtube, or a voice over in the slide file, or shared in a similarly appropriate manner.
   2. The presentation will outline the text used, the problem statement, prior expectations and any insights identified
   3. Slide presentation uploaded to canvas
   4. R script, markdown or notebook supporting the creation of any visuals, models or insights made during the presentation.

Essentially all supporting material including scripts, documents, visuals and/or presentation slides will need to be turned in for review.

## Classes

**Tasks in BOLD are considered assignments and/or extra credit.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | 8-9pm | 9-10pm | Reading Due | **Assignments** |
| Jan 28 | Administrative & Introductions | What is Text Mining? | NA | 1. **Skillset Assessment**    1. Install R/Rstudio on your laptop, or create R Studio Cloud Account    2. Connect to Git Student Repository |
| Feb 4 | R Data Types: Strings | String Manipulation | Chapter 1 | **1.** **Piazza introduction post** |
| Feb 11 | PreProcessing Steps for Text | Term Frequency & Bag of Words | Chapter 2 |  |
| Feb 18 | Associations & Dendrograms | Word Clouds | Chapter 3 | **Extra Credit**  **2. TBD.R** |
| Feb 25 | Polarization | Sentiment Analysis | Chapter 4 |  |
| Mar 4 | ggplot | Ggplot2 | NA | **3. CASE I. Fan Engagement NBA Tweets** |
| Mar 11 | OpenNLP NER | UDPipe: multi-language & lemmatization | Chapter 8 |  |
| Mar 18 | Spring Break | | | |
| Mar 25 | Clustering | Clustering | Chapter 5 |  |
| Apr 1 | LDA | Text2Vec |  | **4. CASE II. Investigative Journalism – Budget Analysis, NER** |
| Apr 8 | Document Classification | Document Classification | Chapter 6 |  |
| Apr 15 | Predictive Modeling | Predictive Modeling | Chapter 7 |  |
| Apr 22 | Gathering Data:  Web-scraping | Gathering Data:  APIs | Chapter  9 |  |
| Apr 29 | Flexdashboard | htmlwidgets |  |  |
| May 6 | Ethics Guest Speaker, *awaiting confirmation or Creating an automated workflow from raw data to dashboard*   * *Ben Taylor, Chief AI Evangelist; DataRobot* * *Mohak Saxena, Data Scientist; DataRoboit* | | | **5. CASE III Text Analysis & Document Classification (graduate students only)** |
| May 14 (Fri) | NA- not a class session | | | **6. Writing Assignment** |

## Graduate Credit Students

This course is open to non-credit, graduate and undergraduate students. As a result, the course experience will vary for each cohort.

Noncredit students may submit case presentations, extra credit homework, and the ethics paper. Your assignments will receive feedback to improve your acumen. However noncredit student may not receive letter grades.

Graduate credit students are expected to do more work and perform at higher standards than undergraduate credit students. As a result, a graduate writing paper needs to demonstrate nuanced sophistication of the ethical considerations presented along with appropriate counter arguments using a philosophical paradigm not opinion. Graduate students must complete all 3 case studies while undergraduate credit students are assigned 3.

## Grading Scale

You earn the grade based on assignments according to the scale below. Grades are not curved to fit a predetermined distribution. A student’s degree, certificate candidacy, or funding status will not have any impact on a course grade. “Needing an A” for any reason is not sufficient to earn an A grade. *Note there are no “minus” grades given in the course.* It is the belief of the instructor that minus grades constitute a false precision in many academic courses and further penalize frequent “A-“ students since there is no way to obtain an “A+” to rebalance a GPA. To the student’s benefit, one can still earn a “plus” on their final grade according to the scale below.

|  |  |  |
| --- | --- | --- |
| Max | Min | Grade |
| 100 | 90 | A |
| 89.9 | 87 | B+ |
| 86.9 | 80 | B |
| 79.9 | 77 | C+ |
| 76.9 | 70 | C |
| 69.9 | 67 | D+ |
| 66.9 | 60 | D |
| 59.9 | 0 | F |