# Course Syllabus: Text Analytics - DRAFT

University of Luxembourg Oct 2021

Dates: Oct 11, 2021 – Oct 22 2021 (specific dates below)

Time: 8:30AM-11AM EST; 2:30pm-5pm CET

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 hosted by TAs

## Important URLs:

Course Moodle: [www.Moodle.uni.lu](http://www.Moodle.uni.lu) (TBD)

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.

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

## Streaming & Video Information:

Lectures will be streamed via zoom.

<https://harvard.zoom.us/j/95561796086?pwd=VDVPMXF6QVNsa2tJZk5TOW5hcys2dz09>

## Prerequisites:

* Textbook: Text Mining in Practice with R ISBN-10: 1119282012
* Software: R & R-Studio
  + This course expects basic understanding of R
  + 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 for interacting during class
* To avoid disruption please install R, R studio and git 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 [www.rstudio.cloud](http://www.rstudio.cloud) but issues may arise due to free tier limits. This is not recommended.

\*\*The professor will try to troubleshoot any installation problems but is not here to perform tier 1 help support. Administrative access, permissions & system settings are limitations not under the control of the professor\*\*

## Course Descriptions & Learning Objectives:

This course is a deep dive into the principles and techniques of text analytics. Topics include text file analysis and construction, reading and writing text files in R, using the APIs for text analysis, and creating frequency histograms for a text corpus and tokens. Students will also

learn how to program in R for effective text analysis. Topics in statistical text analysis will provide working examples and exercises.

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 mathematics is needed but the vast majority of the course content will be focused on applying popular text mining methods. Students will be able to think systematically about how information can be obtained from diverse natural language. The instruction will be case study based with text from various areas of research including journalism, public & governmental interactions, social media, and web sources among others. The course will cover processing text, building visualizations, sentiment analysis and constructing machine learning models along with data ingestion, APIs and web scraping.

Students will learn how to implement a variety of popular text mining algorithms in R.

## Course Learning Objectives

* You will be able to think systematically about how language can be processed and analyzed quantitatively. This objective will be accomplished using ideas from statistics, machine learning and computer science.
* Students will learn how to implement a variety of popular natural language processing methods in R (a free and open-source software) to tackle research problems.
* As a researcher, you will acquire the skill of applying data science concepts within natural language processing to improve outcomes and extract insights.

## Attendance:

Regular attendance (expressed by attending remote sessions) and participation (expressed by interacting in class) is essential to the successful completion of this course. You are responsible for material covered in class even if you have not attended class or watched the recorded lectures. 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 participants, skipping videos and not participating may impact both your assignment sophistication and also your participation grade. 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 University code of conduct. Further, all assignments are the responsibility of each individual pupil unless assigned as a group assignment. Utilizing the peers and the 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 University policies on 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.

## Accessibility

Your professor and the University are committed to providing an accessible, safe, diverse academic community. If necessary, contact school administration for academic, classroom or other appropriate accommodations.

## Grading:

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

* Homework assignments are due the before the start of the next class.
* The case study is due November 5 by 5pm EST (11pm CET)
* The writing paper is due November 5 by 5pm EST (11pm CET)

Assignments are accepted up to 24 hours late with a one letter grade deduction. Any work submitted 24hours will automatically be assigned an F. 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.

Graduate Student Grading

1. Homework I - 5% Basics of R Coding
2. Homework II – 5% Load & Clean documents, identify the most frequent terms
3. Homework III – 10% - Apply several sentiment analyses & cluster a document collection with an unsupervised machine learning approach
4. Homework IV – 10% - Following SEMMA workflow create a document classification model
5. Class Participation – 10% Class participation is not free credit. If students do not contribute, they will not receive class participation credit.
6. Ethics paper Essay – 20% *See below*
7. Case Study – 40% *See below*

## 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. 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 students without significant professional experience, this essay may demonstrate introspection of how society is shaped by this technology and its possible missteps.

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

* Is using a text model to predict candidate expertise in resumes acceptable to save time and money recruiting or does it reinforce historical hiring patterns that are biased?
* Is the technology behind smart speakers that employ natural language analysis helpful or intrusive? Would you as a professional want to work on a project to create this technology in other areas such as smart speakers in the workplace? If these devices are placed in a workplace could they result in liability or bad actors listening in?

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. However, at a miminum tracking sources in a bulleted list at the end of the document is ncessary, simply URLs and book titles with authors should suffice.

### Paper Rubric

|  |  |
| --- | --- |
| Criteria | Reflective Question |
| **Organization** of content– Logical ordering of ideas and focus | Was the paper well organized? |
| **Proofing** – Correct grammar & usage that is appropriate for audience; suitable business and graduate level English language usage | Was the content delivered clearly? |
| **Documentation & Support** – Statements of fact documented, and logically supported | Was no more than 25% of in class articles used? Is there a logical argument made as a personal framework? |
| **Philosophical/Ethical Perspective–** Primary source philosophical references per syllabus and in class lecture to support the perspectives | Did the essay mention or bring in known primary philosophical and ethical frameworks? |
| **Business Perspective–** Recognize the business intersection with ethics? Is the business use case involve data | Was there a business focus in addition to ethics? Did the business examples include non-lecture material? |
| **Broad sophistication-** Demonstration of opposing viewpoints and counter arguments. Effort made to address and overcome these obstacles. | Opposing viewpoints considered? Drawbacks to primary source philosophical frameworks considered? |

## Case Presentations

Each student will work on a case studies individually. Cases will involve using text to apply various methods and draw out insights and conclusions. Each case will have the following work artifacts:

* Maximum 10min voice recorded slide presentation uploaded to youtube, or a voice over in the slide file, screenshare i.e. loom.com or shared in a similarly appropriate manner.
* The presentation will describe and explore text data, the problem statement, prior expectations and any insights identified
* Slide presentation uploaded to the appropriate system
* 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. Late assignments are accepted up to 24hrs late with a 1 letter grade deduction. Assignments submitted more than 24hrs after the due date will automatically be assigned an F.

### Case Rubric

|  |  |
| --- | --- |
| Criteria | Reflective Question |
| **Organization** of content– Logical ordering of ideas, artifacts and visualizations | Was the presentation well organized? |
| **Delivery** – Correct grammar & usage that is appropriate for audience; suitable volume, pace, enthusiasm, posture, and eye contact | Was the content delivered clearly? |
| **Documentation** – proper data support for insights, recommendation, or conclusion with accompanying visual aids | Was the data used? Was it relevant? Did it support the conclusion? |
| **Completeness –** Understood business impact, and mined the data for insights/recommendations | Was the data mined in a significant manner or only cursory? |
| **Data mining Process –** Recognize the type of data mining problem, adherence to established *main* data mining steps. | Did the group approach the problem as follows?  1. Define purpose of proj.  2. Obtain data  3. EDA  4. Partition  5. Method ID & application  6. Interpret, insight, recommendation or implementation |

Class Session Plan

*\*Depending on cohort understanding and dialog, changes are likely to occur to improve learning outcomes.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | If remote:  2:30-5pm Luxembourg  (8:30am-11am EST) | | Reading Due | **Assignments Due** |
| Oct 11  Monday | Administrative & Introductions  What is Text Mining? | R Data Types: Strings | Chapter 1 | Administrative Setup:   * Install R/Rstudio on your laptop, or create R Studio Cloud Account * Connect to Git Student Repository |
| Oct 12  Tuesday | PreProcessing Steps for Text Analysis | Term Frequency & Bag of Words | Chapter 2 | HW1 – Basics of R Coding |
| Oct 13 Wednesday | Associations & Dendrograms, word cloud | Comparison, commonality clouds, word networks, pyramid plots | Chapter 3 |  |
| Oct 14 Thursday | Polarization | Sentiment Analysis | Chapter 4 | HW2 – Load & clean documents, Identify the most frequent terms  *Moved due to HW assignment issue* |
| Oct 15 Friday | OpenNLP NER | UDPipe: multi-language & lemmatization | Chapter  8 |  |
|  | | | | |
| Oct 18 Monday | Clustering | Clustering | Chapter 5 |  |
| Oct 19 Tuesday | Document Classification | Text2Vec | Chapter 6 |  |
| Oct 20 Wednesday | Document Classification - LSA | Predictive Modeling | Chapter 7 | HW3 – apply several sentiment analyses & cluster a document collection with an unsupervised machine learning method  *Moved due to HW new assignment* |
| Oct 21 Thursday | Predictive Modeling | APIs, Webscraping | Chapter 9 |  |
| Oct 22 Friday | Data Science Ethics | Modeling Bias | Ethics Articles | HW4- Following the SEMMA workflow create a document classification model due Oct 23  *Moved due to HW data missing* |

## 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. If the University uses a different scale, i.e. point system, the equivalent percentage grade will be applied to the specific University scale instead of letter grades.

|  |  |  |
| --- | --- | --- |
| Max | Min | Grade |
| 100 | 90 | A |
| 86.9 | 80 | B |
| 76.9 | 70 | C |
| 66.9 | 60 | D |
| 59.9 | 0 | F |