# Course Syllabus: Text Analytics Intensive - DRAFT

## Hult University Spring 2021

Dates: Jan 20-23 2022

Time: 1am-4pm EST Thur; 10am-5pm EST Fri-Sun

Building: NA, Remote only.

Instructor: Ted Kwartler, MBA

Email: edwardkwartler@fas.harvard.edu

Office Hrs: Available upon request

## Important URLs:

* Canvas (homework submissions and grading)

<https://mycourses.hult.edu/courses/3342229>

* The Github repository allows you to get all scripts, PowerPoints and data sets throughout course. 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/Hult_NLP_student_intensive>

### Streaming & Video Information:

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

### 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 di-rectly from the Internet
* A webcam for interacting during class
* 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 instal-lation which may be challenging on Mac OS. As a backup you may use www.rstudio.cloud but issues may arise due to free tier limits.

## 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 mathematical theory 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 in-formation can be obtained from diverse natural language.

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

* CLO 1: Learn how to transform and mine a text file
* CLO 2: Access and use text APIs
* CLO 3: Import, review, manipulate and summarize text data sets in a Term Document Matrix
* CLO 4: Build a statistical model based on a text corpus

### Attendance:

Regular attendance (expressed by watching videos live or asynchronously if the University per-mits) and remote participation (expressed by interacting in class) is essential to the successful com-pletion of this course. You are responsible for material covered in class even if you have not at-tended class or watched 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 Hult University code of conduct. Further, all assignments are the responsibility of each individual pupil unless assigned as a group assignment. Utilizing the forums, online Q/A resources, teaching staff, and/or 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 Hult 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 ex-cuses. There are no excuses for failure to uphold academic integrity.

### Accessibility

Your professor and Hult University are committed to providing an accessible, safe, diverse aca-demic community. If necessary, contact school administration for academic, classroom or other ap-propriate accommodations.

### Grading:

A course grade will be assigned on the basis of student performance on one text based exploratory, visuals, and methods assignment and one case study requiring EDA, visuals and a supervised model. Each assignment is graded out of 100 points and weighted according to the below information.

Assignments are due at 5pm on the data specified in the class table below.

Assignments are accepted up to 24 hours late with a one letter grade deduction. Any work submit-ted 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. A2: NLP Visualizations & EDA 20%: EDA, Methods and Visual Comparisons among two copora
2. A1:   NLP Case – Analyzing Venture Capital Press Releases Case 80%: Press Release Venture Capital Case Study

### Assignment Presentations

Assignment information is contained in the course repository.

Although this is an analytical course, both assignments require presentations to a fictitious non-technical business leader. It is often the case that analytical professionals must demonstrate fluency, earn trust and articulate technical material effectively. Thus, both assignments require the script(s) to produce the technical outputs and a presentation.

Assignments involve using text to apply various methods and draw out insights and conclusions.

The NLP Visualization assignment requires all scripts and a powerpoint of explanations, conclusions or insights with fewer than 15 slides.

**Additionally, the case study requires a narration, or screen share for the presentation.**

The case will have the following work artifacts:

1. 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.
2. The presentation will describe and explore data, 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. Like all assignments, the cases are due at 5pm on the data in the classes table below. Late assignments are accepted up to 24hrs late with a 1 letter grade de-duction. Assignments submitted more than 24hrs after the due date will automatically be assigned an F.

## Classes

|  |  |  |
| --- | --- | --- |
| **Class** | **Covered in Class.** | **Due 5pm EST** |
| Jan 20 Thur | What is NLP, git, r syntax, r-studio |  |
| Jan 21 Fri | String Manipulation, DTM/TDM/WFM, Visuals |  |
| Jan 22 Sat | Supervised & Unsupervised Methods | EDA Visuals Assignment |
| Jan 23 Sun | API, web scraping, ESG & Ethics |  |
| Jan 31 Mon | NA | CASE: Press Releases |

## Grading Scale

You earn the grade based on assignments according to the scale below. 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. Assignments are graded according to a 0-100 point scale with rubrics included on the course site.