# Course Syllabus: Text Analytics - DRAFT

Hult University Spring 2021

Dates: Jan 11, 2021 – Jan 27, 2021 (specific dates below)

Time: 5:00pm-8:00pm

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:

**Canvas** (homework submissions and grading)

TBD

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

TBD

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.

<https://github.com/kwartler/hult_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

1. Software: R & R-Studio
   1. This course expects basic understanding of R
   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.
2. Access to git software to download data sets and class material or ability to download directly from the Internet
3. A webcam for interacting during class
4. 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 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 information 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) and remote participation (expressed by interacting in class and on the class forum) 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 Hult University code of 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 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 excuses. There are no excuses for failure to uphold academic integrity.

Accessibility  
Your professor and Hult 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.

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

Assignments are accepted up to 48 hours late with a one letter grade deduction. Any work submitted 48hours 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. Skills Assessment 5%: **See class info table & student repo** for more information
2. Homework I 10%: **Text Organization**; complete and turn in an R script, notebook or markdown performing the tasks shared in class. Code is graded for accuracy, completeness, logical construction, appropriate comments and being error free.
3. Case I 20%: **Fan Engagement NBA Tweets**
4. Homework II 10%: **Text Document Classification**; complete and turn in an R script, notebook or markdown performing the tasks shared in class. Code is graded for accuracy, completeness, logical construction, appropriate comments and being error free.
5. Case II 30% **E-Sports Marketing – Call of Duty League Team & Player Analysis**
6. Written assignment 15%: **Personal Code of Ethics** essay described below.
7. Class Participation 10%: *Class participation is not free credit. If students do not contribute, they will not receive class participation credit.*

## 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 in 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 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.

## Case Presentations

Each student will work on 2 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:

1. Case I:
   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.
2. Case II:
   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. *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 deduction. Assignments submitted more than 24hrs after the due date will automatically be assigned an F.*

## Classes

**Tasks in BOLD are assignments.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Date | 5-6pm | 6-7pm | 7-8pm | Reading Due | **Assignments Due 5pm** |
| 1 | Jan 11 | Administrative & Introductions | What is Text Mining? | R Data Basics: Files/Repos | Chapter 1 |  |
| 2 | Jan 12 | String Manipulation | PreProcessing Steps to make a Corpus | Term Frequency & Bag of Words | Chapter 2 | **1. Skillset Assessment**   * 1. Connect to Git Student Repository (1pts)   2. Install all packages from the readme (1pt)   3. Complete the basic script in the git repo (2pts)   Students will submit the completed script along with a screen shot of the IDE “Files” pane showing the repo folders. |
| 3 | Jan 13 | Associations & Dendrograms | Word Clouds | Ggplot2 | Chapter 3 | **2. Homework I: Load, clean, make a corpus and DTM** |
|  | Jan 18 | MLK Day Not a Class! | | | |  |
| 4 | Jan 20 | Polarization | Sentiment Analysis I | Sentiment Analysis II | Chapter 4, 6 | **3. CASE I. NBA Fan Engagement** |
| 5 | ~~Jan 21~~ | ~~Clustering LDA & Kmean~~ | ~~Clustering Kmediod~~ | ~~Clustering Spherical Kmeans~~ | ~~Chapter~~  ~~5~~ |  |
| 6 | Jan 25 | Document Classification ElasticNet | Document Classification Bayesian | Document Classification - rtexttools |  | 4. **CASE II. E-Sports Digital Marketing Call of Duty League Team & Player Analysis** |
| 7 | Jan 27 | 3 NBA & 3 Call of Duty Case Student Presentations  Guest Evaluators:  Victor Arias, Managing Account Supervisor, Ketchum Digital: Call of Duty League Marketing  TJ Gable, former Product Marketing Nike | Data Sources & APIs | Ethics, Responsibility & Diversity in analytics and TM |  |  |
| 8 | Jan 30 | Not a Class! | | |  | **5. Homework II: Make a Document Classification Model and Evaluate it**  **6. Writing Assignment** |

## 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.

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