# GSERM Course Syllabus: Remote Text Mining & Natural Language Processing

Times:

9am-12pm; 1-3pm

Instructor: Ted Kwartler

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

Office Hrs: Available upon request

## Important URLs:

**Github Repo:**

## <https://github.com/kwartler/GSERM_ICPSR>

## Prerequisites:

* Textbook: Data Mining for Business Analytics: Concepts, Techniques, and Applications in R

ISBN-10: 1118879368

* 1. Software: R & R-Studio
     1. If you are not familiar with R Studio please take a short introduction to R online 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

## Course Learning Objectives:

If you stay engaged in the course and complete the suggested readings and assignments:

You will be able to think systematically about how language can be ingested and analyzed quantitatively. This objective will be accomplished through the use of ideas from statistics, machine learning and computer science. The pedagogy will be case study based with text from various areas of research including journalism, public & governmental interactions, social media and web sources among others.

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. This course will help introduce the basics of R in text mining but due to course length the topics are not exhaustive.

**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 is essential to the successful completion of this course. You are responsible for material covered in class even if you have not attended class. Given the amount of information covered and short duration, missing one or more class sessions for any reason may result in an automatic reduction in course grade. Unsatisfactory attendance may result in a failing grade. You should plan on spending at least one hour of independent study to prepare for each class session.

## Code of conduct:

This course expects you to uphold and report violations of the University’s code of Academic Integrity. It is your responsibility to know the policy and adhere to it. Cheating on individually assigned reading, exercises, lab work or the final examination or case is considered a violation of the University code of conduct.

You are responsible for understanding University policies on academic integrity. 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  
The University is committed to providing an accessible academic community. Please reach out to the appropriate administration members or your professor should you require academic accessibility support.

## Grading:

A course grade will be assigned based on student performance on four “homework challenges” reviewing the prior day’s lecture, and a text mining case study requiring both PowerPoint and a code submission.

All assignments are needed to earn a satisfactory grade or academic mark. Failure to turn assignments will result in a failing mark.

* Tuesday homework challenge 20%
* Weds homework challenge 20%
* Thurs homework challenge 20%
* Friday homework challenge 20%
* NLP Case Study Script and Slides 20%

## Case Study

The details of the case will be shared at the conclusion of the teaching week. The structure of the case will be one complete R script along with a power point of no more than 10 slides. The entire case study accounts for 20% of the final grade and is evaluated on:

## **Organization of content**– Logical ordering of ideas, modeling artifacts, applicable visualizations in slides – 25pts

## **Organization of code**- R Code is well organized, commented, concise, and free from error – 25pts

## **Data mining Process** – Recognize the type of data mining problem, adherence to established main data and text mining steps. – 25pts

## **Completeness** – Understood business impact, and mined the data for relevant insights/recommendations – 25pts

## Grading Scale

You earn the grade based on assignments above. 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” or equivalent numeric grade (6) for any reason is not sufficient to earn a specific outcome.

The basic schedule is below. However, deviations and changes will occur to improve learning. Further there is an optional lab on specific days for students to bring individual questions related to their specific work, or other inquiries.

|  |  |  |  |
| --- | --- | --- | --- |
| Day | Topics | | Homework Challenges\* |
| Monday | Administration Introduction to R  Setup/Basics | String Manipulation & Text Organization |  |
| Tuesday | Text mining cleaning & basic visualization | Flexdashboard, htmlwidgets | Basics of R Coding |
| Wednesday | Sentiment Analysis & Unsupervised Methods | | Load and clean documents, identify the most frequent words |
| Thursday | Supervised Machine Learning – Document Classification | | Apply several sentiment analyses  Perform 1 clustering method |
| Friday | Ethics, Syntactic Parsing | Robust scripting within the text mining workflow | Following the SEMMA workflow create a document classification model |

*\* Details will be shared in class*