# Course Syllabus: Text Analytics - DRAFT

ICPSR Summer 2024

Dates: May 22,23, 24

Time: 9:30AM-12 EST 12:45-3:45pm

Instructor: Ted Kwartler, MBA

Email:

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

Office Hrs: Available upon request

## Important URLs:

The Github repository allows you to get all scripts, PowerPoints and data sets throughout the 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.

[github repository](https://github.com/kwartler/ICPSR)

## Streaming & Video Information:

Lectures will be streamed via zoom. Please check canvas.

## 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
* To avoid disruption please install R, R studio and git on your local laptop.
* On the last day we will exlpore large language models (LLMs). To do so, students are encouraged to install the software “lm studio” and the Gemma 2B model. Students that are unable to do so, may install “Ollama”. Additional time on day 2 will be devoted to this additional setup.

\*\*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:

Course topics include principles and techniques of text analysis in R. Natural Language Processing (NLP) extracts insights from large amounts of natural language. Lessons will help students add NLP techniques to their research, business and data science toolset.

Curriculum is focused on applying popular text mining methods versus in-depth linguistic theory. Various texts are provided but students are encouraged to work with their own. Technical concepts include processing text, visualization, APIs, web-scraping, sentiment analysis and training ML.

## 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 and participation (if applicable) is essential to the successful completion of this course. You are responsible for material covered in class even if you have not attended class and are participating remotely. Missing more than 1 class session for any reason may result in an automatic reduction in course grade (if applicable) or denial of an attendance certificate. You should plan on spending at least one to two hours of independent study outside of class

## Code of conduct:

This course expects you to uphold and report violations 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

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:

As a short course no ssessments will be utilized. However, to support a portfolio artifact or current research an optional Thursday lab will be offered allowing students to apply methods to their own text with direct instructor support.

Class Session Plan

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

|  |  |  |
| --- | --- | --- |
| **Class** | **Lesson Topics** | **Optional Assignments & Recommended Reading Due** |
| May 22 Weds | * What is NLP, git, r syntax, r-studio? * Preprocessing steps: string manipulation, term frequency | Chapter 1 |
| May 23 Thu | * Bag of Words DTM/TDM * Visualizations – wordclouds, histograms, pyramid plots, word networks, dendrograms, associations, dendrograms | “Homework” HW1-Basics of R Coding  Chapter 2, 3 |
| May 24 Fri | * Sentiment Basics * Basic Supervised & Unsupervised Methods: Classification & Clustering * GPT: How do LLMs work? Prompting Strategies & Agentic Workflows | “Homework” HW2 Load, Clean documents & Frequency |
| May 24 Fri  **4-6pm** | OPTIONAL LAB: Bring your own data & ideas to work alongside the professor! | |