# GSERM Course Syllabus: Text Mining & Natural Language Processing

University of Ljubljana

Dates: Aug 26, 2019 to Aug 30, 2019

Times:

9:15 – 11:45pm Morning Class

1:30 – 3:15pm Afternoon Class

Building: TBD

Instructor: Ted Kwartler, MBA

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_TextMining>

## Prerequisites:

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

ISBN-10: 1118879368

* 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
* Access to git software to download data sets and class material or ability to download directly from the Internet
* Students are encouraged to sign up at <https://rstudio.cloud/> for a hosted R-Studio instance. Since each pupil’s personal computer differs, the professor will only provide technical support for R-Studio’s cloud instance. As a result, please sign up for a free account as a back up to your personal laptop.

## 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 found [here](https://www.uni-lj.si/university/code_of_ethics/). Cheating on individually assigned reading, exercises, lab work or the final examination 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 on the basis of student performance on the final (80%) and a written assignment (20%). Pupils will attend a class session for in person final exam proctoring.

During the final exam, no phones, tablets or computers should be used even as calculators. If you need a calculator you must bring one to your examination period. A student may prepare a single, double sided 3inch by 5inch, *handwritten* index card for use during any examination. Cards that are larger, typed or multiple cards will constitute cheating according to the University’s academic integrity policies.

*For most students a simple calculator may be needed but keep in mind the exam seeks to determine conceptual understanding and sophistication not mathematical ability. Thus some students may not need a calculator at all.*

* Final Exam 80% of final grade
* Written assignment 20% of final grade

## Writing Assignment

The writing assignment will be due 2 weeks after class is completed. No late submissions will be accepted.

Twenty percent of the final grade will be determined by the quality and completeness of a 900 to 1200 word ***essay concerning ethical implications of text mining***. Approximately, no more than 25% of the essay should comprise a summary and synthesis of the assigned data science ethics articles. The balance of the essay can incorporate new literary sources and student reflections for how ethics can play an informative role for data science and text mining applications.

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

* Is it ok that Amazon’s Alexa has humans listening to speech to improve the algorithm even if most users do not know this?
* Should public text (facebook pages) be analyzed for marketing (Facebook ads), insurance (searching for risk factors), or polictical targeting (Cambridge Analytica)?
* Is there an ethical duty to tell users you are collecting information and reselling it or simply bury it in a terms of service agreement? Does anyone really read the agreements?
* Is text mining medical journals an appropriate method for deriving diagnosis and treatment plans? What if this enables underserved populations to gain some level of healthcare expertise?

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 text mining.

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.

## 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.