**Natural Language Processing with Bag of Words & LLM Methods**

**Prerequisites (knowledge of topic)**This course assumes no prior experience with machine learning or R, though it may be helpful to be familiar with introductory statistics and programming.

**Hardware**A laptop computer is required to complete the in-class exercises.

**Software**R (https://www.r-project.org/) and R Studio (https://www.rstudio.com/products/rstudio/) are available at no cost and are needed for this course.

**Learning objectives**Students will understand how to extract, interpret, and measure content from natural language data, utilizing popular NLP methods to identify key insights and emotional content.

**Course content**  
The Introduction to Natural Language Processing (NLP) at GSERM is a comprehensive journey into the world of textual data analysis. The course is designed to immerse attendees in both the theory and practical implementation of versatile NLP methods, transforming qualitative research prospects.  
Through a mix of lectures and labs, participants will gain practical proficiency in powerful NLP techniques that include:

* Large Language Models (LLMs)
* Prompt Engineering
* Vector Database Basics
* Bag-of-words Analysis
* Sentiment Analysis
* Document Classification and Clustering

Students with previous experience in programming, graduate-level statistics, and mathematical theory will benefit most from this course. However, the curriculum is crafted to appeal and be accessible to all researchers eager to integrate NLP tools in their analysis.

**Structure**The course effectively blends theoretical knowledge and hands-on experience. Your typical day in the course will include:

* Morning Session: Engaging lectures and demonstrations on a specific NLP technique using the R language.
* Afternoon Session: Practical application sessions where the technique of the day is applied to new datasets.
* Lab: An opportunity to apply the day’s learned theory and practice on a provided dataset, or students can choose to use their own. Instructor support will be available throughout these lab hours.
* This approach ensures participants gain a clear theoretical understanding of NLP, as well as the practical ability to implement text mining techniques with confidence.

*Day 1: R Basics & Introduction to NLP*

* Intro to R programming
* Introduction to NLP & basic text mining
* String Manipulation & Text Cleaning

Lab Section: Clean tweets, and prepare for bag of words examination  
  
*Day 2: Visualizations in text mining*

* Word Frequency & Term Frequency Inverse Document Frequency (TF-IDF)
* Term Document, & Document Term Matrices
* Word Clouds – Comparison Clouds, Commonality Clouds
* Other Visuals – Word Networks, Associations, Pyramid Plots, Treemaps

Lab Section: Create various visualizations with news articles

*Day 3: Sentiment Analysis & Machine Learning: Document Classification & Clustering*

* Lexicon based sentiment analysis
* Elastic Net (Lasso & Ridge Regression)
* K-Means, K-Mediods & Spherical K-Means

Lab Section: Classify clickbait from news headlines, group news articles by clusters

*Day 4: Introduction to Large Language Models*

* How do LLMs work?
* Accessing LLMs with APIs, libraries gptstudio/gpttools/ chattr, or local LLMs
* Common LLM tasks: Sentiment, document classification, Named Entity Recognition, summarization, POS tagging

Lab Section: Using an LLMs to classify clickbait from news headlines, group news articles by clusters and compare to Day 3’s lab results

*Day 5: Effective Prompt Engineering, vector databases and RAG models*

* Introduction to prompt engineering for effective LLM usage
* Introduction to Vector Databases
* Building a basic RAG LLM workflow in R for information retrieval

Afternoon Session: Bring your own research data so we can explore it in a lab setting!

**Literature**

**Mandatory**

* Text Mining in Practice with R by Ted Kwartler; Wiley & Sons Publishing

ISBN: 978-1-119-28201-3

* Two Data Ethics articles assigned at class to spur reflection for the ethics essay.

**Supplementary / voluntary**

None.

**Mandatory readings before course start**

* Read chapter 1 of Text Mining in Practive with R entitled “What is Text Mining?”
* Please install R & R Studio on your laptop prior to the 1st class. Be sure that these are working correctly and that external packages can be installed. As a backup, sign up for an account at R-Studio’s cloud environment [https://rstudio.cloud](https://rstudio.cloud/).

**Examination part**

20% Ethics Paper – *Due at midnight at the last course day*

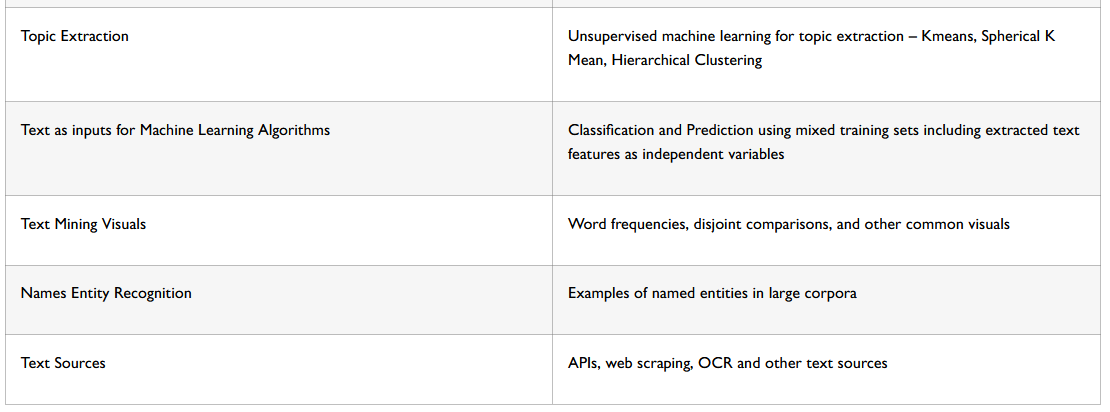
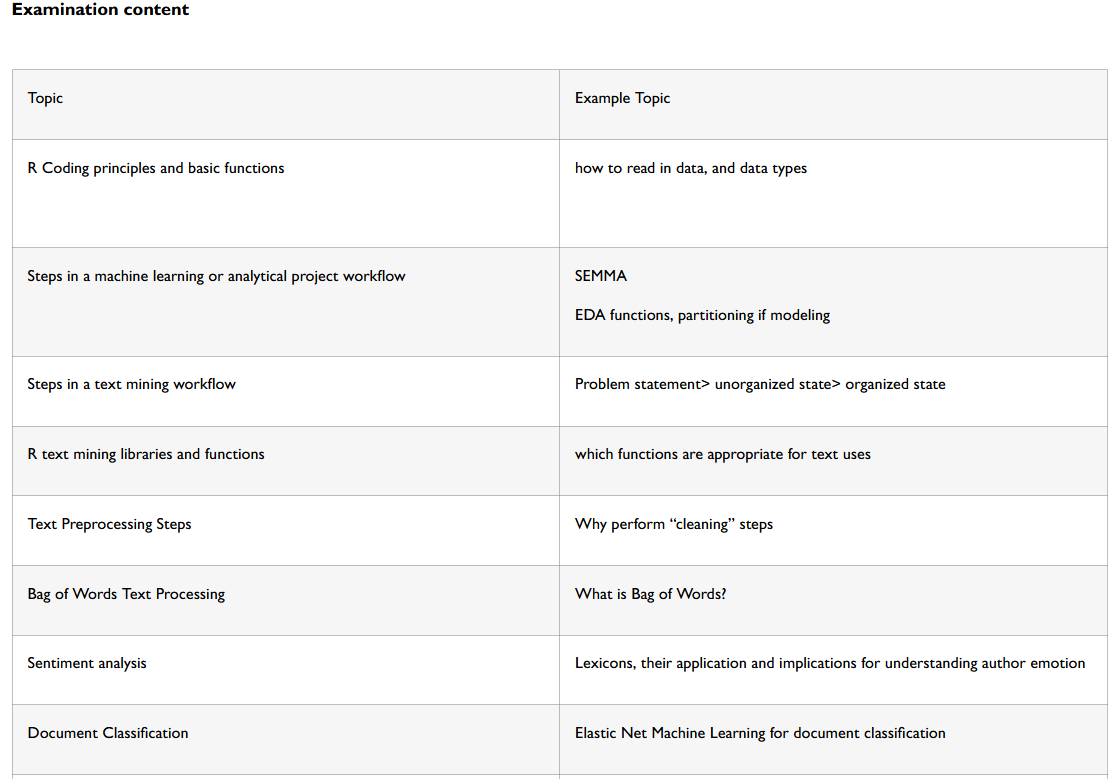
* 500-750 word essay with personal reflection on the ethical implications of text mining research methods

80% Final Exam – *Proctored on the final day of the week*

* 30 multiple choice (2pts each),
* 1 of (20pts) code review section asking students to describe *what* and *why* specific code steps are being taken
* 4 of (5pts each) Short form questions/answers requiring 1 paragraph (2-4 sentences each)

**Supplementary aids**

Students may bring a hand written “index card” to the final examination period. It may be double sided, and should be functionally equivalent to the UK standard 3in by 5in notecard. Students may put any information they deem important for the final on their notecard and use it as a supplement during the exam. Use of an exam supporting notecard is optional.



**Literature**

The exam will be based on the lectures and mandatory assigned reading from Text Mining in Practice with R.