Doctoral School in Finance and Economics

Course ID Course Title
Introduction to Natural Language Processing

1. Course details

Semester:	1
Credit rating:	2ECTS/30 TU
Pre-requisite(s):	Graduate statistics, some programming knowledge in any language recommended not required
Lecturer(s):	Prof Kwartler
Administrator:	Roswitha Glorieux
Tutor(s):	
Seminar times and rooms:	Winter semester 2021/2022
Tutorial times and rooms:	Hybrid (depending covid evolution) dates TBD Oct 11-15 in person if possible (5hrs per day) Oct 11-15 & 18-22 if remote (2.5hrs per day)
Communications	**It is important that students should regularly read their University e-mails, as important information will normally be communicated this way.**
Mode of assessment:	Case study, Ethics essay and homeworks
Examination Periods:	Case study due 2 weeks after the last class Homework assignments due before the start of the next class Ethics Essay due 2 weeks after the last class
Course WebPage:	Moodle.uni.lu

2. Aims and objectives

Aims
The instruction will be case study based with text from various areas of research including journalism, public & governmental interactions, social media, and web sources among others. The course will cover processing text, building visualizations, sentiment analysis and constructing machine learning models along with data ingestion, APIs and web scraping.
Learning Objectives
1. 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. 2. 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. 3. As a researcher, you will acquire the skill of applying data science concepts within natural language processing to improve outcomes and extract insights.

3. Plan of semester

Date	If remote: 2:30-5pm Luxembourg (8:30am-11am EST)		Reading Due	**Assignments Due**
Oct 11 Monday	Administrative & Introductions What is Text Mining?	R Data Types: Strings	Chapter 1	Administrative Setup: • Install R/Rstudio on your laptop, or create R Studio Cloud Account • Connect to Git Student Repository
Oct 12 Tuesday	PreProcessing Steps for Text Analysis	Term Frequency & Bag of Words	Chapter 2	HW1 – Basics of R Coding
Oct 13 Wednesday	Associations & Dendrograms, word cloud	Comparison, commonality clouds, word networks, pyramid plots	Chapter 3	HW2 – Load & clean documents, Identify the most frequent terms
Oct 14 Thursday	Polarization	Sentiment Analysis	Chapter 4	
Oct 15 Friday	OpenNLP NER	UDPipe: multi-language & lemmatization	Chapter 8	
Oct 18 Monday	Clustering	Clustering	Chapter 5	HW3 – apply several sentiment analyses & cluster a document collection with an unsupervised machine learning method

Oct 19 Tuesday	Document Classification	Text2Vec	Chapter 6	
Oct 20 Wednesday	Document Classification - LSA	Predictive Modeling	Chapter 7	
Oct 21 Thursday	Predictive Modeling	APIs, Webscraping	Chapter 9	HW4- Following the SEMMA workflow create a document classification model
Oct 22 Friday	Data Science Ethics	Modeling Bias	Ethics Articles	

4. Course details (by topics)

The course will first introduce the statistical language R, the R-Studio integrated development environment (IDE) and basic data types including "strings". Additionally a basic review of communication theory, the challenges of text analysis and an overview of the upcoming topics are shared.

Once this foundation is established, functions to manipulate and "clean" string data types will be covered. Ultimately the frequency count of tokens/terms within strings and documents will organized into a document term matrix (DTM), thus beginning the bag of words style analysis methods.

After organization into a DTM, methods such association, frequency visualizations, dendrograms and comparisons between document collections can be performed. Next, a linguist theory "Zipf's Law" will help contextualize the use of various sentiment analysis lexicons.

Lastly, basics of machine learning, particulary centered on NLP applications will be explored. These include unsupervised methods to extract document clusters, topics and prototypical documents as well as classification methods within supervised machine learning models.

5. Reference list/ Bibliography

Text Mining in Practice with R by Ted Kwartler (the instructor)
https://www.amazon.com/Text-Mining-Practice-Ted-Kwartler/dp/1119282012
ISBN-13: 978-1119282013
ISBN-10: 1119282012

Johns Hopkins Berman Institute of Bioethics: IBM Pitched It's Watson Supercomputer as a Revolution in Cancer Care
http://bioethicsbulletin.org/archive/ibm-pitched-its-watson-supercomputer-as-a-revolution-in-cancer-care

In 2016, Microsfot's Racist Chatbot Revealed the Dangers of Online Conversation
https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

An LSA Package for R
Fridolin Wild, Vienna University of Economics and Business Administration, Austria
http://nm.wu-wien.ac.at/research/publications/b675.pdf

Spherical K-Means Clustering

Journal of Statistical Software
https://pdfs.semanticscholar.org/90a4/e7af0698a7c3a82a67d55557a4bf9b012c4b.pdf?_ga=2.60679470.208021855.1622225402-504860289.1622225402

Building the State of the Art in Sentiment Analysis of Tweets
National Research Council Canada
http://www.cs.toronto.edu/~xzhu/SEMEVAL_13.pdf

6. Further information about assessment

Mode of assessment:	• Natural language case study including code for construction of all analytical artifacts, a powerpoint explaining the concepts and insights identified and an accompanying video or narration of the slides.
	• 1200 word ethics essay regarding the use of NLP technology, which may cover one of the following topics
	  ○ Using NLP for resume reviews, efficiency gain or biased? How could any bias be addressed or is it completely unacceptable?
	  ○ Use of smart speaker technology, is it helpful, manipulative, or a violation of privacy? What about smart speakers as a form of surveillance in the workplace where employees are unable to object?
	  ○ Are NLP models biased against non-English as a frist language speakers? How could this affect someone's lives if they have an accent, speak with different socio-economic related lexicons? How could this be protected against?
	• HW1 – Basics of R Coding
	• HW2 – Load & clean documents, Identify the most frequent terms
	• HW3 – apply several sentiment analyses & cluster a document collection with an unsupervised machine learning method
	• HW4- Following the SEMMA workflow create a document classification model

Examination Periods:	Case study due 2 weeks after the last class Homework assignments due before the start of the next class Ethics Essay due 2 weeks after the last class
Rating:	Homework 30% marks, Class participation 10%, Essay 20%, Case study 40%.

Ethics Paper Rubric

Criteria	Reflective Question
Organization of content– Logical ordering of ideas and focus	Was the paper well organized?
Proofing – Correct grammar & usage that is appropriate for audience; suitable business and graduate level English language usage	Was the content delivered clearly?
Documentation & Support – Statements of fact documented, and logically supported	Was no more than 25% of in class articles used? Is there a logical argument made as a personal framework?
Philosophical/Ethical Perspective– Primary source philosophical references per syllabus and in class lecture to support the perspectives	Did the essay mention or bring in known primary philosophical and ethical frameworks?

Business Perspective– Recognize the business intersection with ethics? Is the business use case involve data	Was there a business focus in addition to ethics? Did the business examples include non-lecture material?
Broad sophistication- Demonstration of opposing viewpoints and counter arguments. Effort made to address and overcome these obstacles.	Opposing viewpoints considered? Drawbacks to primary source philosophical frameworks considered?

Case Study Rubric

Criteria	Reflective Question
Organization of content– Logical ordering of ideas, artifacts and visualizations	Was the presentation well organized?
Delivery – Correct grammar & usage that is appropriate for audience; suitable volume, pace, enthusiasm, posture, and eye contact	Was the content delivered clearly?
Documentation – proper data support for insights, recommendation, or conclusion with accompanying visual aids	Was the data used? Was it relevant? Did it support the conclusion?
Completeness – Understood business impact, and mined the data for insights/recommendations	Was the data mined in a significant manner or only cursory?
Data mining Process – Recognize the type of data mining problem, adherence to established *main* data mining steps.	Did the group approach the problem as follows? 1. Define purpose of proj. 2. Obtain data 3. EDA 4. Partition 5. Method ID & application 6. Interpret, insight, recommendation or implementation