Advanced Quantitative Text Analysis

PHD COURSE, SAMPOL903

Dr. Nicole Baerg

March 28-29, 2022

UNIVERSITY OF BERGEN

Course content

This two-day workshop will cover natural language processing and quantitative text analysis using the R statistical environment and quanteda. We will also cover companion packages for text analysis, namely spacyr for part of speech tagging and dependency parsing. We will cover some intermediate text topics and the course assume familiarity with basic text analysis topics. If you are a bit rusty, I highly suggest you visit the quanteda website and have a look at the tutorials. The course will be very hands on, with lots of time to do applications in predesigned notebooks.

Learning outcomes

Upon successful completion of this course participants should be able to:

- Understand some of the most common machine learning methods used for solving classification problems
- Provide an overview of recent research linking text analysis and causal inference
- Explain different models of text
- Extract quantitative information from text and analyze this information using R

Course programme

Total Lecture: 8 hours
Total Lab: 4 hours

Day 1 Morning: 9-12

Introduction (or reminder) to supervised Machine Learning models

- Naïve Bayes
- Support Vector Machines

Lecture and examples: 9-11
Lab exercise 11-12

Day 1 Afternoon: 13-16

Causal inference for Machine Learning

- Text as omitted variable
- Text and matching

Lecture and examples: 13-15
Lab exercise 15-16

Day 2 Morning: 10-12

Different Models of Text:

- Bag of Words vs Contextualized Topic Models
- Embeddings and Latent Sematic Analysis

Lecture and examples: 9-11
Lab exercise 11-12

Day 2 Afternoon: 13-16

Information Extraction and NLP:

- Named Entity Recognition
- Other IE tasks

Lecture and examples: 13-15
Lab exercise 15-16