Statistics 653: Statistical Natural Language Processing (2 units)

Spring 2021

Prof. Suess, Department of Statistics and Biostatistics, CSU East Bay

Lecture:

- Section 1: MW 8 to 9:40, SSc 146, Zoom

Intructor: Prof. Eric A. Suess **Office:** NSc 319 **Phone:** 510-885-3879 **e-mail:** eric.suess@csueastbay.edu

Office Hours: MW 2 to 3pm, or by appointment

Class Website: http://www.sci.csueastbay.edu/~esuess/stat653

Required Text:

- Julia Selge, David Robinson, Text Mining with R, A Tidy Approach, O'Reilly, 2019.
- tidy-text-mining

Reference Texts:

- Baumer, Kaplan, Horton, Modern Data Science with R, CRC Press, 2017.
- Lantz, Machine Learning in R, Second Edition, Packt, 2015.
- quanteda
- Natural Language Processing with Python
- gensim

Further References:

- Hadley Wickham and Garrett Grolemund, R for Data Science, O'Reilly 2018.

Material To Be Covered:

This is the fourth course in the sequence of Data Science courses offered by the Department of Statistics and Biostatistics for the M.S. Data Science Concentration. The Data Science courses are specifically for registered students in the M.S. Statistics program.

The sequence of courses are:

1. Stat. 650 Advanced R for Data Science
2. Stat. 651 Data Visualization
3. Stat. 652 Statistical Learning
4. Stat. 653 Statistical Natural Language Processing
5. Stat. 654 Introduction to Applied Deep Learning

These courses are intended to be taken in order as they build upon each other, but you can discuss taking the courses out of order with instructor approval.

The topics of the course will follow the topics presented in the Modern Data Science with R book. The book will be used as the primary text for Statistics 650, 651, 652, 653. For each class there will be other supporting reference materials.

The main topics for Statistics 652: Statistical Learning

- Chapter 1 Tidy Text Format
- Chapter 2 Sentiment Analysis
- Chapter 3 Word and Document Frequency
- Chapter 4 Relationships between words
- Chapter 6 Topic Modeling

Homework: A list will be on the website. Homework will be assigned weekly. Homework will be "due" on Mondays, which means you should complete the homework and come to class prepared to ask questions. Homework will be "collected" though Blackboard and needs to be submitted by Wednesday of the week the homework is due.

Quizzes and Exams: Two short quizzes, one midterm will be given and the final.

Grading:

- Project 30%
- Homework 15%
- Quizzes 5%
- Midterm 25%
- Final 25%

Policy on Make-up Exams: You are expected to take the quizzes and exams at the scheduled times. In case of genuine emergency, illness or hardship, for which you can present written documentation I may agree to arrange for a make-up exam. Make-up exams must always be arranged BEFORE the regular exam is given and always take place AFTER the regular exam. Quizzes may not be made up!

Statistics 652 SLOs

Student Learning Outcomes (SLO's):

Students graduating with an M.S. in Statistics from Cal State East Bay will be able to:

1. Apply statistical methodologies, including a) descriptive statistics and graphical displays, b) probability models for uncertainty, stochastic processes, and distribution theory, c) hypothesis testing and confidence intervals, d) ANOVA and regression models (including linear, and multiple linear) and analysis of residuals from models and trends at the Master's level.
2. Derive basic theory underlying these methodologies.
3. Model practical problems for solutions using these methodologies.
4. Produce relevant computer output using standard statistical software and interpret the results appropriately.
5. Communicate statistical concepts and analytical results clearly and appropriately to others; and,
6. Employ theory, concepts, and terminology at a level that supports lifelong learning of related methodologies.