Stefan Eng

Education

- *Master's Programme in Mathematical Statistics at Gothenburg University.
- Bachelor of Science in Mathematics & Computer Science at California State University, Northridge

Work Experience

Scientific Computing @ Novartis

June 2019 - August 2019

- 3 month internship in the Scientific Computing and Consulting (SCC) group within Biostatistics at Novartis in Basel, Switzerland.
- Developed a TTE (time-to-event/survival analysis) R shiny application for oncological clinical trial data with support for subgroup and subpopulation analysis.
- Created a patient subpopulation creation and ad-hoc listing R shiny app for medical writers and clinicians to faster query specific patients from CDISC-ADaM data sets.
- Analyzed results of protocol amendments of clinical trials.

Programmer/Analyst @ CRESST/UCLA

August 2016 - June 2018

- Worked on research projects involving natural language processing (NLP)
  - Presented results at IEEE Conference on Semantic Computing Eng, Tan, and Iseli (2018)
- Created an interactive game using Angular2/Ionic to test and measure reasoning and problem solving where students attempted to solve tasks by programming an on-screen vehicle.
  - Conducted multiple data collections at high schools across Los Angeles involving around 100 students.
  - Event data was analyzed and exploratory analysis was conducted to summarize student's problem solving and reasoning abilities.

Data Scientist/Software Developer @ JPL

June 2013 - May 2016

- Managed data processing pipeline for Cassini and Dawn missions which includes Scala/Apache Spark ingest code and managing Elasticsearch cluster.
- Wrote parsers to bring old proprietary data formats into common and modern formats. Including XML parsing, web scraping, and Scala's parser combinators.
- Lead a project to extract data, using Apache Tika, from word documents into MongoDB and presented with a web interface.

Projects

subpat

subpat is a collection of R shiny modules to create subpopulations and subgroups of clinical trial data. It was designed with CDISC ADaM data format in mind but supports any data format. It features two main applications, the Patient Listing Generator (PLG), and a TTE (time-to-event) analysis app.

https://github.com/Novartis/subpat

This was developed during my summer 2019 internship at Novartis at the Scientific Computing and Consulting group.

1990 to 1992 Census Crime Rate Prediction

- Date: 2019-2-25
- Blog post: https://stefanengineering.com/1990-1992-census-crime-prediction
- Code: https://github.com/stefaneng/Census-Crime-Rate-Prediction

In this analysis I investigate a dataset that provides some demographic information for 440 of the most populous counties in the United States in years 1990-92. First I explore linear regression models and then a negative binomial regression model to predict the number of crimes per 1000 people in each county. The best linear regression model performed similarly to the negative binomial regression model on the training and test set using the same variables.

Cloze Deletion Prediction with LSTM Neural Networks with Keras

- Date: 2018-11-5
- Blog post: https://stefanengineering.com/cloze-deletion-prediction
- Code: https://github.com/stefaneng/Cloze-Deletion-Prediction

A cloze deletion test is a form of language test where a sentence (or paragraph) is given to the test taker with blanks for missing words. The student is expected to fill in a "correct" word in the blanks.

I explore various machine learning approaches to fill in these missing words from sentences from two Swedish news sources. The goals for this paper were to answer the following questions:

- Can we predict missing word using only the words around it?
- What sentences are good example sentences?
- Does length of sentence make a difference?
- Where are good sources to find cloze deletion sentences?

I compare the difference between an LSTM (Long-Short term memory) neural network with that of a Bidirectional LSTM using Keras using python.

Awards

- Received Staff Appreciation and Recognition (STAR) award 2018 at UCLA
- First place at HackPoly Hackathon Spring 2016
- First place in CSUN's Computer Science Senior Design Project Showcase 2016
- First Place in 2014 SS12 Code for a Cause Competition, a programming competition to create applications for people with disabilities.

Conferences

[1] S. Eng, J. Tan, and M. Iseli. "Adjective Intensity Ordering by Representing Word Definitions as a System of Linear Equations". In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC). IEEE, Jan. 2018. DOI: 10.1109/icsc.2018.00047. URL: https://doi.org/10.1109/icsc.2018.00047.