DEREK B. YEN

400 West 45th Street, New York, NY · derekyen@nyu.edu · +1 (858) 699-9684
Website: dbyen.github.io · LinkedIn: linkedin.com/in/derek-yen

EDUCATION

New York University Expected graduation: May 2020

- Master of Science in Data Science (GPA: 3.96)
- *Current and past coursework:*
  - Deep Learning for Natural Language Processing, Machine Learning, Optimization and Computational Linear Algebra, Probabilistic Time Series Modeling

University of California, Los Angeles June 2018

- Bachelor of Science in Mathematics of Computation with Minor in Neuroscience
- *Selected coursework:*
  - Information and Power, Algorithms in Bioinformatics, Mathematical Modeling

SELECTED PROJECTS

- *News Article Text Classification*: Compared sentiment analysis methods (Naive Bayes, bag-of-words support vector machines, and neural networks) for set of articles from The New York Times
- *Recommender system with PySpark*: Used ALS for implicit feedback in PySpark on Last.fm dataset. Experimented on cold starting using a latent feature regression model for unknown items
- *Building a Semantic Parser to Handle Queries about Song Data*: Applied the SEMPRE framework to a parser which can answer queries about a subset of the Million Song Dataset
- *Numerical Linear Algebra*: Implemented SoftImpute-ALS matrix completion algorithm in Python for recommendations with testing on MovieLens-100K
- *Capstone: Applying a General Language Model to Medical Text*: Applied multilabel text classification strategies to dataset of 6M+ NYU Langone Health medical notes with deep NLP models in PyTorch

SKILLS AND QUALIFICATIONS

Programming languages:	Python, R, MATLAB, C++, HTML, JavaScript
Python library proficiencies:	PyTorch, Scikit-Learn, Pandas, NumPy, SciPy, Tensorflow
Big data tools:	SQL, PySpark, MapReduce
Design and typesetting:	LaTeX, WordPress, Adobe InDesign

WORK EXPERIENCE

NBCUniversal Media

Intern for Data Science Practica with Corporate Decision Sciences Jan. 2019 - Dec. 2019

- *Statistical testing and analysis*: Conducted statistical testing to measure differences in age group distributions over time, under both independence and autocorrelation assumptions. Built visualization tool to simulate reallocation across the distribution using Retool and JavaScript.
- *Evaluating forecasting models*: Developed processes using MLFlow and R to save 800+ forecast models and evaluate based on MAPE, visualization, etc. Designed application for senior data scientists to track model performance (folded into the production process)
- *Developing forecasting models*: Produced 18-month forecasts with Python machine learning (ARIMA, regression, random forest) for shortform digital content including YouTube, Hulu, etc.

GuidedChoice

Customer data analytics intern for retirement planning financial services firm Sept. 2017 - Dec. 2017

- Produced insightful visualizations using Python and demographic breakdowns for Florida account

JOURNALISM AND LEADERSHIP EXPERIENCE

The Daily Bruin

Copy chief at award-winning UCLA student newspaper June 2016 - June 2017

- Managed scheduling, payroll, and development of the Daily Bruin style guide
- Responsible for ensuring content is edited for print (7K copies/day) and web (400K views/month)
- Led weekly meetings with seven slot editors and monthly training sessions with 30 contributors

RESEARCH EXPERIENCE

UCLA Mathematics Department

Student research assistant modeling dynamics of homeless populations in LA Sept. 2017 - July 2018

- Extracted and cleaned data from public census databases and geocoding APIs using Python
- Analyzed and visualized city regions with principal component analysis, nonnegative matrix factorization, correlation analysis, etc.
- Implemented and evaluated machine learning models such as logistic regression with Tensorflow for predicting large changes in homeless populations
- Communicated results to UCLA faculty and students through multiple presentations and reports

Swartz Center for Computational Neuroscience at University of California, San Diego

Research assistant for experiments on neural correlates of competition June 2015 - December 2017

- Wrote Python scripts to collect data on subject responses
- Developed MATLAB code for preprocessing EEG data using EEGLAB software
- Supported grant applications with detailed plots and statistical trend information