DANIEL FREES

Data Scientist

@ dfrees@stanford.edu ☎ (650) 336-3260 📍 SF Bay Area in linkedin.com/in/danielfrees github.com/danielfrees

EXPERIENCE

Data Scientist

IBM

📅 June 2022–Present 📍 San Francisco, CA

- Develop custom predictive and generative models using NLP, traditional ML algorithms, and AI engineering of LLMs
- *Example 1:* Developed a python package and trained NRE & RE models to extract graphical representations of insurance claims documents and store this data in neo4j
- *Example 2:* Engineered software which enables automated generative summarization of massive financial documents such as 10ks

Research Data Scientist

UCLA BrainSport Lab

📅 May 2020–June 2022 📍 Los Angeles, CA

- Lead research to develop a machine learning model which identifies brain injuries in rsfMRI data `https://github.com/danielfrees/braingraphML`
- Designed data analysis software which enabled the discovery of new trends in rsfMRI research methodology (Paper submitted to publications Oct 2023). `https://github.com/danielfrees/rsfMRI_LitReview`

NLP Engineer

UCLA Loes Lab

📅 Nov 2020–May 2021 📍 Los Angeles, CA

- Designed and tested context model for determining affirmation/negation of medical conditions in electronic health records (EHR)

OPEN-SOURCE PROJECTS

ScrapeMed

NLP Data Scraper for PubMed Central 📅 Sep 2023

- Designed, developed and distributed an open-source Python package for scraping and cleaning data from PubMed Central.
- Notable features: natural language paper search, pandas integration, XML validation, data reference mapping, interactive PMC search
- Duke University is using ScrapeMed to power surgeryGPT

`https://github.com/danielfrees/scrapemed`

RAC++

High-Speed, Scalable Hierarchical Clustering 📅 June 2023

- Co-developed a python-wrapped C++ algorithm for faster agglomerative clustering of large datasets.
- RAC++ is over 100x faster than sklearn's Agglomerative Clustering model at 35,000 pts, and can support tasks with hundreds of thousands of datapoints where sklearn times out

`https://github.com/porterehunley/RACplusplus`

EDUCATION

Stanford University
2023-2025

MS Statistics and Data Science

University of California, Los Angeles
2018-2022

BS Computational Biology: Data Science

GPA: 4.0

Relevant Coursework

Machine Learning, Data Science, Data Management Systems, Artificial Intelligence, Algorithms & Complexity, Data Structures & Algorithms (C++), Intro Computer Science (C++), Systems Modeling, Digital Image Processing, Computer Organization, Bioinformatics, Probability & Statistics, Linear Algebra & Applications, Discrete Math, Differential Equations

SKILLS

- Python
- Data Science
- Software Dev
- Prompt Engineering
- Linux
- Data Cleaning
- NLP
- C++
- Leadership
- Consulting
- IBM Cloud
- LaTeX
- BioInformatics
- SQL
- NoSQL
- HTML & XML
- Graph Theory

AWARDS

- IBM 2023 Growth Award
- IBM L2 Certified Expert Data Scientist
- UCLA Summa Cum Laude Honors
- UCLA Alumni Scholars Top Scholarship
- UCLA Dean's Honor List (12x)
- Frank Livermore Trust Scholarship
- Eagle Scout Award
- CCS Scholar Athlete of the Year