Zhihao **Du**

zhihao617@berkeley.edu | +1 (510) 833-4417 | GitHub | Personal Website

Education

M.Anlytx (Analytics) – **University of California, Berkeley**   08/2023 - 05/2024
GPA: in progress   Berkeley, CA

B.A. Computer Science, Statistics – **University of California, Berkeley**   08/2019 - 05/2023
GPA: 3.8/4.0   Berkeley, CA

Skills

Languages – Python, SQL, Java, C, R, C#, HTML+CSS.

Coursework – Natural Language Processing, Machine/Deep Learning, Statistical Learning, Service Operations Design, Database Systems, Reproducible Data Science, General Linear Models, Data Structures, Machine Structures, Linear Algebra, Probability & Statistics.

Tools/Frameworks – Python Frameworks: Numpy/Pandas, Matplotlib, Scikitlearn, Pytorch/Jax, Flask. MS Azure, MySQL Database, MongoDB, .NET, DBeaver, Git, Unix System, Microsoft Office.

Experiences

Research Assistance – **Project AEI**   01/2022 - 05/2023
Real-time audio emotion classification   Berkeley, CA
- Implemented the Conformer neural network using PyTorch, then experimented, trained, and finetuned the model on MS Azure platform to classify streaming real-life audio speech data into seven universal emotions;
- Spearheaded and pipelined training data preprocessing with speech augmentation techniques including Gaussian white noise, simulated room impulse response, and randomly sampled background noise; Boosted testing accuracy from 68% to 71% and ensured consistent evaluation performace on real-time audio inputs;
- Used pyaudio and librosa to program and installed audio streaming and continuous model evaluation on a Raspberry Pi 4 device by downsizing audio sample rate and performing batched evaluation to accomodate for the resource-limited platform;

Data Engineer Intern – **DataCVG Co Ltd**   05/2021 - 08/2021
Relational database design and extract-transform-load pipelines with SQL   Shanghai, China
- Performed extract-transform-load (ETL) on client FosunPharma's two major relational database, resolved conflicts with hand designed schemas, and consolidated all data into a single database;
- Programmed 100+ complex SQL query scripts and engineered data merger pipelines using the DBeaver ETL software;
- Overcame data confidentiality through extensive communication with client representatives and increased client's database system throughput by 80%, benefiting downstream business intelligence tasks.

Technical Projects

Howamidoing Full stack web application development with Flask   01/2023 - present
- Designed a college level course grade tracker and class standing estimator web application using the Flask framework, HTML, CSS, and JavaScript, in addition to the truncated normal probability distribution to model student grades;
- Implemented and optimized JSONizable user data objects and stored in NoSQL database with connection to MongoDB;
- Utilized the g variable to store the user object for easy I/O accross multiple different requests;

Zilean Python package for data mining and engineering pipelines   05/2022 - 08/2022
- Developed python package "zilean" that bridges the Riot Games API with traditional python data science APIs (scikitlearn, pandas) to produce data pipelines for multidimensional data ready for downstream ML or DL tasks;
- Programmed, tested and refined data mining/engineering algorithms for large semi-structured with rate limiting API request algorithms;
- Promoted and published as open source project with immediate collaborators after established CI/CD pipelines using Github Actions and Readthedocs documentation

Leadership Experiences

Teaching Assistant – **University of California, Berkeley**   01/2023 - 05/2023
CS182/282A: Deep Neural Networks   Berkeley, CA
- Led, facilitated and supported students on weekly discussion sections, homework parties, and project office hours;
- Built and distributed a interactive demo on BERT using Jupyter Notebook through designing a surrogate task that quickly showcased the effectiveness of pretraining on BERT with self-supervised learning;