YEWEN ZHOU

https://www.linkedin.com/in/yewen-zhou/ (626) 492-9028
https://hegelim.github.io/personalwebsite/ yz4175@columbia.edu

EDUCATION

Columbia University New York, NY
M.S. in Data Science Dec 2022

- GPA: 3.80 / 4.0
- Coursework: Algorithms, Big Data, Causal Inference, Machine Learning, Applied Deep Learning (fall 2022), Finance for DS

University of California, Berkeley Berkeley, CA
B.A. in Data Science, Business Analytics Concentration May 2021

- GPA: 3.93 / 4.0, Phi Beta Kappa Society, High Distinction Honor
- Coursework: Data Structures, Time Series, Artificial Intelligence, Probability and Statistics, Decision Analytics, Intro to Finance

SKILLS & TECHNOLOGIES

Programming:	Python, Linux, SQL, R, Java, C++, HTML5, CSS, JavaScript
Python Packages:	pandas, pytorch, tensorflow, keras, pyspark, numpy, scipy, scikit-learn, matplotlib
Frontend Frameworks & Cloud Services:	Django, Bootstrap, Plotly, AWS, Google Cloud Platform
Development Tools:	Git, Jupyter, Docker, VSCode, PyCharm, IntelliJ, RStudio
Writeup:	Markdown, reStructuredText, LaTeX

WORK EXPERIENCE

Scry Analytics, Inc San Jose, CA
Data Science and Engineering Intern May 2022 – Aug 2022

- Benchmarked 30 text recognition models from 5 open-source repositories using PyTorch, Docker, AWS
- Generated synthetic dataset from 1,791 images with existing tags for chart detection model training
- Reduced ABINet recognition model inference time by half, significantly making the current product more competitive
- Trained detectron2 deep learning model for chart detection with image augmentation, achieving 82 AP in test set
- Contributed to a million-dollar worth project in extracting key-value pairs from bar charts and finished the base version

iQIYI, Inc Beijing, CN
Ads Algorithm Backend Intern May 2021 – Aug 2021

- Developed a testing framework for ads allocation emulator with more than 10,000 records; deployed in the server launched overseas in more than 5 countries (pandas, logging, numpy)
- Created a SARIMA time series module for ads inventory prediction, achieving a cross-validation RMSE less than 0.2 (statsmodels)
- Implemented High Water Mark (HWM) algorithm from scratch (logging, numpy, pandas) based on Yahoo research paper for compact allocation; used as the 1st version by algorithm and product teams of more than 10 people

PROJECTS

Realtime Twitter Sentiment Analysis (Python) Nov 2021 – Dec 2021

- Developed 6 ML models including Linear Regression, Ridge Regression, Gradient Boosting, AdaBoost, Random Forest, and SVR for aggregated twitter sentiment prediction, attaining test RMSEs less than 0.1 (sklearn)
- Leveraged Virtual Machine (VM) on Google Cloud Platform (GCP) to decrease model training time by 16x
- Created a dashboard using Bootstrap, Django, HTML5/CSS/JavaScript/Plotly, displaying real-time Twitter sentiment prediction

Stock Price Prediction (Python) Nov 2021 – Dec 2021

- Utilized Airflow Scheduler to collect stock prices from 5 tech companies automatically daily at 7 am
- Trained and updated 5 linear regression models for stock price prediction; obtained relative errors less than 0.01

The solveminmax Python Package (Python) Jul 2021 – Sep 2021

- Implemented an object-oriented, open-source Python module to solve a sum of min and max equations (regular expression, numpy, scipy, matplotlib)
- Designed unit tests (pytest) to validate module extensively with more than 30 testing cases
- Distributed on the Python Package Index (PyPI) with documentation on GitHub written in reStructuredText and Markdown

Gitlet (Java) Apr 2020 – May 2020

- Built a version-control system that mimics the features of Git using data structures, OOP, graph-traversal, file-reading/writing, serialization, and error-handling

Ordering Menu (C++) Oct 2018 – Dec 2018

- Implemented a real-world online-ordering menu that allowed the user to select different items and accessories
- Improved the practicality of the program by implementing combos, gifts based on user's choice
- Generated a complete ordering receipt based on the user's choice, payment method, and personal information