YEWEN ZHOU

https://yewenzhouofficial.com (626) 492-9028
https://www.linkedin.com/in/yewen-zhou/ thefirstzyw@gmail.com

EDUCATION

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

- GPA: 3.83 / 4.0
- Coursework: Algorithms, Big Data, Causal Inference, Machine Learning, Applied Deep Learning, Finance for DS

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

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

SKILLS & TECHNOLOGIES

Languages:	SQL, Python, Jupyter, Linux, 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, Docker, VSCode, PyCharm, Rstudio

WORK EXPERIENCE

Scry Analytics, Inc San Jose, CA
Data Scientist Mar 2023 – Present

- Researched 30 Large Language Models (LLMs) to pinpoint prime integration candidates for existing products
- Led a team of 5 Data Analysts and trained an Arabic text recognition model with 1M dataset, achieving 0.98 test accuracy
- Designed an Arabic synthetic data generator, generating a comprehensive dataset of 450k samples for Arabic detection training
- Enhanced business capabilities by integrating a chart-extraction pipeline into existing products through Docker containerization

JPMorgan Chase New York, NY
Data Science Intern Sep 2022 – Dec 2022

- Developed heatmaps for data visualization comparing land cover change distributions across four categories from 2013 to 2017
- Constructed U-Net and Fully Convolutional Network (FCN) models using Tensorflow Keras for land cover change prediction, achieving 0.4 mean Intersection over Union (IoU)

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

- Led a $1M project focused on key-value pair extraction from bar charts, fostering client engagement and future partnerships
- Benchmarked 30 text recognition models from 5 open-source repositories using Pytorch, Docker, AWS
- Generated synthetic dataset from 1,791 images with existing tags, optimizing chart detection model training
- Halved ABINet recognition model inference time, significantly boosting product competitiveness
- Trained detectron2 deep learning model for chart detection, achieving 82 Average Precision (AP) in the test set

Columbia University SAFE Lab New York, NY
Data Scientist Sep 2021 – May 2022

- Conducted ETL data analysis on a 50 GB child-abuse database using SQL queries, delivering valuable insights
- Enhanced workload efficiency by innovatively deploying AWS Athena for SQL execution on large datasets

PROJECTS

Columbia University, Detecting Cancer Metastases on Gigapixel Pathology Images Sep 2022 – Dec 2022

- Curated 16,439 patches from 9 gigapixel images for a training dataset exceeding 30 GB, driving high-quality model development
- Developed customized Deep Learning models based on InceptionV3 that concatenates images at 2 zoom levels
- Achieved 0.8 AUC, 0.84 accuracy, and 0.75 recall detecting tumors in 1 gigapixel test image at 2 zoom levels

Columbia University, StackOverflow Data Analysis Feb 2022 – Mar 2022

- Conducted detailed user behavior analysis on over 18 million StackOverflow entries utilizing advanced SQL techniques
- Employed Google Cloud Platform's BigQuery for expedited processing of large datasets

Columbia University, Realtime Twitter Sentiment Analysis Nov 2021 – Dec 2021

- Created 6 accurate ML models (Linear Regression, Ridge Regression, Gradient Boosting, AdaBoost, Random Forest, and SVR) with Scikit-learn for Twitter sentiment prediction, achieving test Root Mean Squared Errors (RMSEs) < 0.1
- Leveraged Google Cloud Platform's Virtual Machines to expedite model training time by a factor of 16
- Developed a real-time Twitter sentiment prediction dashboard using Bootstrap, Django, HTML5, CSS, JavaScript, and Plotly