Mengyu Zhang

Seattle, Washington, United States
6264384723 | sam.zhang.069@gmail.com
LinkedIn | Github

EDUCATION

Johns Hopkins University, Master of Science, Data Science — May. 2023

UCLA Extension, Data Science Certificate Program — Mar. 2021

University of California, Los Angeles, Bachelor of Science, Applied Mathematics Major, Statistics Minor — Dec. 2020

- Honors: Dean's Honor List

PROFESSIONAL EXPERIENCE

Data Science Intern — California, U.S.

Sikka Software Corporation — May. 2022 – Aug. 2022

- Built **ETL Pipelines** with **PySpark** and **Boto3** to update and maintenance over 1 billion patients' record stored on **AWS S3**.
- Connected **AWS Athena** with **PyAthena** and **SQLAlchemy** to implement data management and preview on collected data.
- Constructed **BiLSTM** model with **PyTorch** on **AWS EC2** to predict health indicators for evaluating patients' dental health conditions.
- Detected medical entities from medical notes data with **AWS Comprehend Medical** to perform **explanatory data analysis** identified insights on time sequential procedure records, demonstrating feasibility in predicting progressive diseases' procedures.
- Performed word embedding for sequential text using **BioBERT** and **multi-task learning** models constructed by **TensorFlow Keras**.

Data Science Intern — Beijing, China

Inspur Group Company — Jun. 2021 – Jul. 2021

- Built **Selenium WebDriver** on **Python** to enable automatic web crawler for collecting and updating over 1 million receipt images' data.
- Designed image editing filters with **OpenCV** to effortlessly change brightness, contrast and add data augmentation to images.
- Developed and deployed **YOLOv4** object detection with **CUDA**, generating prediction on receipts' recognition led to 95%+ accuracy.
- Built **OCR** invoice recognition product with **Pytesseract** and achieved 77.4% OCR accuracy, an improvement over 30% baseline.
- Built **PowerBI Dashboards** to visualize and present the product evaluation results to the senior management team.

Data Analyst Intern

BroadStreet Data Co-operative — Sep. 2020 – Dec. 2020

- Developed interactive **Tableau dashboard** with ability to visualize COVID data by location and real-time analysis.
- Conducted statistical analysis (fourth order Runge-Kutta) to investigate infection rate and exposure from Los Angeles' historical data, identified insights on attributes for building dynamic SEIR model to simulate separation of COVID-19.
- Engaged in depth-analysis of industry advancements through scientific journals and self-guided academic exploration.
- Provided ad hoc support to VP in assessing the feasibility of proposals to maximize efficiency by enabling strategic planning to increase technical allocation, assigning top-tier technicians to high-priority projects, ensuring clients' satisfaction.

PROJECT EXPERIENCE

Machine Learning in Institution Name Disambiguation from Scholarly Text | IEEE Capstone Project — Dec. 2021 – Jun. 2022

- Implemented **Python** script with **Psycopg** to update database for IEEE's data lake of over 5.4 million papers in **AWS Redshift**.
- Performed **Named Entity Recognition** with **spaCy,** created geolocator with **GeoPy** to impute geographical features.
- Applied **LightGBM Classifier** with Python to perform supervised pairwise comparison and classification.
- Deployed **Hierarchical Agglomerative Clustering (HAC)** and Research Organization Registry API to center and standardize clusters.
- Published algorithms and data with **Sphinx** documentation on **AWS EC2**, achieved 81.4% accuracy.

Using NLP and Language Models on Financial Documents to Predict Stock Price Movement — Mar. 2022 – May. 2022

- Built **Web Scraping Pipeline** with **Beautiful Soup** and **Requests** to fetch stock prices and 8-K files data from Yahoo and Morningstar.
- Pre-processed texts with **NLTK WordNet** corpus reader in combination with **Dask** for multithreading speedboost.
- Implemented **AllenNLP** to build **ELMo** model for news texts' tokenization, part of speech tagging and sentences segmentation.
- Performed word embedding for sequential text using **BERT** and **FinBERT** models with **PyTorch**.
- Designed different deep learning models (**MLP**, **RNN**, **LSTM**) with **TensorFlow Keras** to predict future prices and achieved 88% recall.

SKILLS

Programming Languages: Python, R, SQL, Java, C++, MATLAB

Tools & Libraries: PySpark, TensorFlow, PyTorch, Keras, spaCy, CoreNLP, scikit-learn, Numpy, Pandas, NLTK, Seaborn, MySQL, SQLite, PostgreSQL, AWS Redshift, AWS S3, AWS EC2, AWS Athena, AWS Comprehend, CUDA, MLFlow, Git, Bash, Docker, VBA, Conda

Core Competency: Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, Data Mining, Data Visualization, Data Analytics, Optimization, Algorithms, Data Structure, Database, Object Oriented Design