Yanwen Lin

Phone: (+1)-412-708-5446
Email: lyw1124278064@gmail.com

Education Background

Carnegie Mellon University
Master of Science in Intelligent Information Science (School of Computer Science), ***GPA****: 3.8/4.0* Aug.2019-Dec.2020 (expected)
Master of Science in Civil and Environmental Engineering, ***GPA****: 3.9/4.00* Aug.2017-Dec.2018
Selected Courses*: Distributed Systems, Parallel Computer Architecture and Programming, Introduction to Deep Learning, Cloud Computing, Search Engines, Computer Networks, Introduction to Computer System*

Dalian University of Technology
Bachelor of Engineering in Civil Engineering, ***GPA****: 3.8/4.00 (top 10%)* Sep.2013 - Jun.2017

Professional Skills

- **Programming Languages**: Java, Python, C, Scala, Bash, HTML/CSS/Javascript, MATLAB, R
- **Software and Tools**: AWS, PyTorch, Numpy, Pandas, MySQL, HBase, MongoDB**,** Hadoop, Spark, Kafka, CUDA, OpenMP, MPI, Java Spring/Spring MVC, Java RMI, Terraform**,** Docker and Kubernetes, Pandas, Git, YACC/Flex

Work Experience

Development on OLAP Big Data Platform, *SWE Backend Intern* Horizon Robotics, Nanjing, Jan.2019-Mar.2019
- Integrated Apache Druid with access control system using basic security and Kerberos extension.
- Coordinated pluggable Apache Kylin with Hadoop computing engine, HBase data storage and Hive data warehouse.

Longitudinal evaluation on Deployed Pittsburgh Fire Risk Model Metro21 Inst., Pittsburgh, Jun.2018-Oct.2018
- Built an evaluating system which feeds ~600k lines of fire and property data within entire Pittsburgh into a XGBoost model to estimate its performance using Python Pandas and Jupyter Notebook.
- Identified potential factors that leads to high fire risk based on model result and informs the Bureau of Fire's prioritization of property fire inspections.

Paper

Jessica Lee, **Yanwen Lin**, Michael Madaio. A longitudinal evaluation of a deployed predictive model of fire risk. 32nd Conference on ***Neural Information Processing Systems*** AI for Social Good Workshop, Montréal, Canada.

Projects

AFS-like Distributed File System Based on Check-on-use Cache Policy Jan.2020-Feb.2020
- Developed a distributed file system including interposition shared library and RPC server from scratch.
- Designed a complicate RPC protocol message format for communication between RPC client and server.
- Implemented open-close session to resolve conflicts of sharing files between concurrent users.
- Integrated check-on-use cache proxy and LRU eviction policy to reduce file retrieval latency.

Parallel Acceleration for Rat Colonies Migration Simulator Feb.2020-Mar.2020
- Implemented two parallel versions of a rat migration simulator using OpenMP and MPI, respectively.
- Achieved significant speedup than sequential simulator (for OpenMP 7x speedup, for MPI 6x speedup, at a 12-core machine)

High Performance Web Service for Data Retrieval Oct.2018-Dec.2018
- Conducted Extract, Transform and Load on a large Tweets dataset (~ 1 TB)
- Developed user intimacy ranking system and topic word extraction system based on pre-processed Twitter data and provided APIs for client queries.
- Optimized various aspects of the system such as database schema, cluster load-balancing, data sharding and replication.
- Achieved 6th in a 6-hour live server-performance competition out of 32 teams.

Stateful Stream Processing for Finding Client-Driver Matching Dec.2018
- Deployed Kafka jobs to produce driver location, event streams and Samza jobs to output client-driver match.
- Employed RockDB as the persistent key-value store to maintain the state during stream processing.
- Managed and monitored Samza jobs on YARN using logging and yarn web-UI utilities.

Iterative Processing System on Social Relationship Graph Data via Spark Oct.2018-Nov.2018
- Developed a ***Spark*** application processing Twitter social graph data (~10 GB).
- Implemented PageRank algorithm using ***Scala*** to rank user by influence.
- Improved model performance via profiling tools from 1 hour to less than 30 minutes given limited computing resources.

Big data Analytics with Large Wikipedia dataset Sep.2018
- Processed Wikipedia dataset (~340 GB) to retrieve daily hot topics via implementing Hadoop jobs deployed on AWS EMR.
- Employed defensive programming and test-driven developing techniques such as Junit, MRUnit.