Yanwen LIN

Phone: +1-412-708-5446 | E-mail: yanwenl@andrew.cmu.edu
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

Education Background

Carnegie Mellon University Aug.2017 - Present
Master of Science in Civil and Environmental Engineering GPA: 3.92/4.00
Dalian University of Technology Sep.2013 - Jun.2017
Bachelor of Engineering in Civil Engineering GPA: 3.84/4.00

Research Experience

Longitudinal evaluation on Deployed Pittsburgh Fire Risk Model Metro21 Inst., Carnegie Mellon University
Research Assistant, Partnered with the City of Pittsburgh's Bureau of Fire Jun.2018-Oct.2018
Key words: Deployed model evaluation, Consistency assessment

- Designed consistency assessment framework for deployed Pittsburgh fire risk model along timeline.
- Conducted longitudinal evaluation on deployed model via various metrics such as transition measurement between different risk-level groups, top-k empirical risk curve.
- Accepted by NIPS AI for Social Group as workshop paper.

Bridge Placement Optimization for world-wide Rural Area Carnegie Mellon University
Research Assistant, Partnered with Bridge to Prosperity Organization Jun.2018-Jul.2018
Key words: Exploratory Data Analysis, Decision making modeling

- Cleaned and explored rural area bridge assessment dataset, including correlation and geo-spatial analysis.
- Designed a mathematical model to quantify the connectedness impact for the rural area.
- Performed global sensitivity analysis on the model parameter using Sobol Indices via R.

Virtual Tracking Experimental System Based on Unity 3D Dalian University of Technology
Traffic Science and Technology Competition work Oct.2014 - Apr.2015
Keywords: Virtual experimental system, Unity 3D

- Developed a virtual tracking experimental system programmed by C# in Unity3D.
- Awarded the first prize in the 10th Traffic Science and Technology Competition in DUT.

Paper

(Accepted) Jessica Lee, **Yanwen Lin**, Michael Madaio. A longitudinal evaluation of a deployed predictive model of fire risk. In Proceedings of the 32th Conference on Neural Information Processing Systems, AI for Social Good Workshop, 2018.

Course Project

High Performance Web Service for Data Retrieval Oct.2018-Dec.2018, CMU
Key words: Integrated system, Undertow, HBase, AWS Elastic MapReduce, Performance tuning, Terraform

- Implemented Extract, Transform and Load (ETL) on a large Tweets dataset (~ 1 TB) and loaded the data into MySQL and HBase systems (with customized MapReduce using Bulk-Loading API).
- Orchestrated frontend Undertow server and backend HBase/MySQL server on cloud infrastructure using Terraform.
- Developed data analysis APIs supported by user intimacy ranking system and topic word extraction system.
- Optimized each piece of the system to improve throughput from ~1000+ to ~8000+ RPS.

Social Networking Timeline with Heterogeneous Backends Oct.2018, CMU
Key words: Database-as-a-Service, Social networking system, MySQL, Neo4j, MongoDB

- Integrated RDBMS (MySQL), GraphDB (Neo4j) and NoSQL (MongoDB) in a social network web service context.
- Built a complex social networking web application with fan-out queries that span multiple databases.

Iterative Processing System on Social Relationship Graph Data via Spark Oct.2018-Nov.2018, CMU

Key words: Apache Spark, Social graph data, Distributed system profiling

- Developed Spark applications progressively using composite of RDDs, DataFrame and SparkSQL.
- Finished the execution model of graph processing by analyzing a social graph with the PageRank algorithm.
- Profiled Spark applications with YARN resource management system and Web UI.

Cannes Film Prediction based on Imbalanced Twitter Data Modeling Apr.2018-May.2018, CMU

Key words: Data science pipeline, Imbalanced dataset, ROC, Confusion matrix

- Developed a pipeline for data collection, pre-processing and modeling to predict 2018 Cannes winner list.
- Tuned model Hyper-parameters based on Receiver Operating Characteristic (ROC) curve and Confusion Matrix.

Applied Machine Learning for Gene Expression Profile Classification Nov.2017-Dec.2017, CMU

Key words: Multilabel classification, Dimension Reduction, Ensemble

- Filtered meaningful features via deep exploratory data analysis and knowledge of the problem domain.
- Performed dimension reduction on cell gene data (~20,000 dimension) using PCA, LDA to maximize variance inside sampled data and speed up data processing
- Constructed a multi-label classification model (ensembled by logistic regression, SVM, Gaussian Process) based on pre-processed training data and produced prediction for types of testing cells.

Professional Skills

- **Programming language**: Python (Pandas, Scikit-learn, Tensorflow), Java (Maven), Bash script, Scala, C, MATLAB, R
- **Frameworks:** Hadoop MapReduce, Apache Spark, Undertow, Vert.x, Flask, Scrapy, Apache Sqoop
- **Cloud computing:** AWS (EMR, ELB, Autoscaling, CloudWatch, Rekognition, SNS), GCP (App Engine, ML Engine, AutoML), Azure (HDInsight), Infrastructure as code with Terraform, Function as Service
- **Database System:** RDBMS (MySQL, SQLite), NoSQL (HBase, MongoDB), GraphDB (Neo4j)
- **Technical skills:** Docker and Kubernetes, Machine learning, Statistical analysis, Git, Yaml, Latex, Mob Programming