Linh Hoang Truong

CCA175 Cloudera Spark and Hadoop Certified Developer

mrthlinh@gmail.com
truonghoanglinh
mrthlinh
4694078033

EDUCATION

University of Texas at Dallas, Richardson, TX Jan 2017 - Dec 2018
MSc Computer Science with concentration on **Data Science**

Danang University of Technology, VN Sept 2010 - June 2015
BSc Electronics and Communication Engineering

SpringBoard Data Science Program Apr 2018 - Oct 2018
6-month intensive program in Data Science

EMPLOYMENT

DATAECONOMY, Data Engineer, Dublin, OH Jan 2019 - Present
- Executed multiple internal / external projects on Cloudera Big Data Distribution software in different use case settings.
- Using HDFS, Sqoop, Python, Spark, Hive, Flume to handle data ingestion, processing and consumption.

Superior Data Science, Data Scientist Intern, Dallas, TX Aug 2018 - Nov 2018
- Developed production level machine learning model to improve accuracy of predicting task completion time in company's flagship project management
- Delivered 8% reduction in number of task delayed over time
- Worked with Flask Framework and Tensorflow Serving to deploy trained model

Erik Jonsson School of Engineering and Computer Science, Research Assistant, Richardson, TX Jan 2017 - Apr 2018
- Applied Deep Learning (Keras and Deeplearning4J) to help software engineers writing code more efficiently, fixing bugs automatically, parameter recommendation, automatic type resolution of variables
- Co-author of two accepted research papers in top-tier conference (acceptance rate around 20%)

Robert Bosch Engineering and Business Solutions, Software Engineer Intern, HCM city, Vietnam Mar 2015 - July 2015
- Integrated Fuzzy Logic and Genetic Algorithm to Car Simulator in order to improve the accuracy of Antilock Braking System and Traction Control System by 5%

SKILLS

Coding: Python, Java, R, Javascript, Scala
Machine Learning: Scikit-learn, Keras, Tensorflow, PySpark, Deeplearning4J
Data Engineer: HDFS, Hadoop, Spark, Hive, Sqoop, Flume
Database: MySQL, MongoDB, HBase
OS: Linux, Window
Cloud Services: AWS, Google Cloud

PROJECTS

Agile User Story Point Estimation (Tensorflow, Python) Aug 2018 - Nov 2018
- Developed an end-end solution (Natural Language Processing) for effort estimation in Agile development without manual feature engineering
- Experimented with Word2Vec, Doc2Vec, Tf-idf for feature extraction and LSTM model for deep learning model.
- Average error of getting correct story point is 1.9, outperform manual feature engineering by 8%
- Report: https://github.com/mrthlinh/Agile-User-Story-Point-Estimation

FIFA World Cup 2018 Winner Prediction (Sklearn, Pandas, Numpy) Apr 2018 - June 2018
- Automatically crawled data from various sources based on designed features.
- Hypothesis testing with two-sampled t-test to evaluate feature importance.
- Experimented with many machine learning algorithms such as logistic regression, SVM, Random Forest, Neural Network, Gradient Boosted Tree
- Predicted win / lose/ draw with 56% accuracy and 33% for goal difference for matches in the tournament.
- Report: https://github.com/mrthlinh/FIFA-World-Cup-Prediction/blob/master/report/report.md

Spotify Playlist Continuation Recommender (Pyspark, Sklearn, Pandas) July 2018 - Oct 2018
- Developed a recommender system for automatic playlist continuation on 66 million ratings dataset.
- Compared performance between various type of recommenders, K-Nearest Neighbor and Collaborative Filtering on Pandas and PySpark
- Increased the chance of users getting their favorite songs in a playlist to 78%
- Report: https://github.com/mrthlinh/Spotify-Playlist-Recommender

Google Analytics Customer Revenue Prediction (Kaggle Project, Sklearn, Pandas, Numpy) Aug 2018 - Nov 2018
- Developed a Machine Learning model to predict how much Google Store customers will spend based on 55 features
- For every customer, model could predict how much they spent with error within $5, top 15% in Leader Board
- Report: https://github.com/saodem74/Google-Analytics-Customer-Revenue-Prediction

RELEVANT COURSES

Statistical Methods for Data Science, Statistical and Machine Learning, Machine Learning, Big Data Management and Analytics, Database Design

PAPERS

- Co-author "Complementing Global and Local Contexts in Representing API Descriptions to Improve API Retrieval Tasks" , *Foundations of Software Engineering (FSE) 2018* - **Acceptance Rate (21%)**
- Co-author "Statistical Learning of API Fully Qualified Names in Code Snippets of Online Forums publication date" , *International Conference on Software Engineering (ICSE) 2018* - **Acceptance Rate (24%)**

CERTIFICATES

CCA175 Cloudera Spark and Hadoop Developer Apr 2019
Deep Learning Specialization (Neural Network, Convolution Neural Network, Sequence Model) - Coursera Dec 2018