Yangtao Zhang

(734) 353-1029 | maxzhang@umich.edu| LinkedIn: Yangtao-Zhang/

EDUCATION

University of Michigan
Ann Arbor, MI

Master of Science, Data Science (GPA**: 3.88/4.0**) Sep.2019 – May.2021

Course Highlights: Machine Learning, Data Mining, Natural Language Processing, Reinforcement Learning, Information Retrieval, Information Visualization, Computational Data Science, Data Manipulation, Database Management

Shanghai Jiao Tong University
Shanghai, China

Bachelor of Engineering, Computer Engineering (GPA: **3.60/4.0**) Sept.2015 – Aug.2019

Course Highlights: Data Structures, Algorithm, Operating System, Linear Algebra, Statistic Methods in Engineering

WORK EXPERIENCES

Data Scientist Intern
Ann Arbor, MI

University of Michigan, Office of University Development, Data Decision & Support June.2020 - Present

- Wrote Python, Spark and SQL for data preparation and analysis, including machine learning modeling, segmentation, and matching on a database of over **2M** potential donors via Databricks on Microsoft Azure cloud platform
- Cleaned multi-source, multi-format data bought from various vendors and crawled from LinkedIn, Zillow by building a compatible data pipeline deployed as automated jobs on Databricks
- Constructed predictive models using **XGBoost** and **Bayesian Network** to forecast the expected donation from donors
- Increased the expected donation by **80%** through recommending potential wealthy donors with collaborative filtering. Presented outcome to the whole data science team and the model has been widely used and favored by gift officers
- Analyzed the factors contributing to the probability of donating and prioritized potential donors based on wealth, inclination, and other features for different units. Conducted **A/B testing** on various group to verify the sensitivity of the analysis. This project was selected and presented to **VP** by my manager and will be applied to the next campaign

Machine Learning Research Assistant
Ann Arbor, MI

University of Michigan, Interaction & Collaboration Research Lab, Sponsored by Dell Dec.2019 - Oct.2020

- Collaborated with a team of data scientists from Dell to research on deep learning models including **BERT, DistillBERT, XLNet, Electra**. Achieved best performance of **0.89** F1-score for 20-category aspect extraction and **0.68** MAE for 9-level sentiment analysis. Completed machine learning pipeline and set up the whole service on ***Dell EIG***'s internal Linux server
- Improved **Electra** with post training on domain knowledge extracted from reviews on Amazon electronics products
- Built a web crawler of Amazon product reviews. Aggregated crowdsourcing data labelled by Amazon MTurk workers via calculating worker-worker agreement, annotation similarity. Cleaned and transformed results to MongoDB Atlas
- Delivered impactful presentations with interactive visualizations to demonstrate how sentiment towards Dell products changed over time, what aspects of Dell products that customers cared most, etc. to further underscore the power of the machine learning models and the business impact to stakeholders and managers.

Software Development Engineer Intern
Shanghai, China

XueYao Education Technology May.2019-Aug.2019

- Worked full stack with React/Redux, Django, MongoDB to create a responsive, multi-user web app, which automated daily manual work conducted by the operation team and reduced their workload from **1 day to 1 min**
- Built an automatic homework grading pipeline which can identify and classify mistakes using spaCy and NLTK, generate adequate recitation material based on mistakes and behavior data, report engagement of students by visualizing learning history, and notify the operation team with emails of students with abnormal behavior data
- Developed REST/HTTP APIs with authentication, supporting customized query via JSON/XML/HTML data formats

Data Engineer Intern
Shanghai, China

GE Appliances Jan.2018-May.2018

- Developed an ETL pipeline in Python and Spark for data collecting, data cleaning, data analysis and data visualization of crowd-funding market to support data-driven decisions for business team with a **10k** daily volume
- Conducted cluster analysis of trending products and keywords in crowd-funding market using **K-means, TF-IDF**
- Delivered an interactive dashboard of trending products' history, real-time popularity and potential future change

SKILLS

Programming Languages:	Python, Java, C++, JavaScript, HTML, CSS
Frameworks/Tools:	Pandas, Scikit Learn, PyTorch, TensorFlow, Tableau, Docker, Kubernetes, Git, AWS
Databases/Big Data:	MySQL, MongoDB, Elasticsearch, Spark, Hadoop

PROJECTS

Machine Learning: Sentiment Analysis with First GOP Debate Tweets

- The project focused on analyzing sentiment of tens of thousands of tweets discussing the first GOP debate and built both ML methods (SVM, Random Forest, Logistic Regression) and DL methods (LSTM, Bi-LSTM) to predict polarity
- Conducted data cleaning by lowering and removing non-English characters. Handled data imbalance with up/down sampling and class weight assignment. Extracted word embeddings using word2vec and pretrained embeddings
- Tuned ML model parameters using grid search and DL model hyperparameters using random sampling on log scale to minimize the range. The final model achieved 0.70 F1-score, which beat the baseline model with 10%