Wenyi (Wesley) Tao

(201) 201-9455 * wt2271@columbia.edu * https://wesleytao.github.io/

WORK EXPERIENCE

02.2019-present **Data Scientist,** Ushur Inc Santa Clara, CA

- Built a reporting data pipeline which hourly log, track and anonymize batch of data from multiple production instance to a centralized data warehouse and performed advanced query in MongoDB, leveraged data visualization and analytics tools to provide KPI survey response rate, completion rate to the client and management.
- Built a statistic component which create hourly, daily, monthly bucket to store a variety of statistics across multiple enterprises accounts in the system and implemented an API which provide statistics for an arbitrary time window
- Developed an NLP (natural language processing) pipeline for insurance underwriter's decision automation process.
- Integrated a dockerized rule-based expert system UMLS (unified medical language system) for feature extraction which greatly improved coverage for disease detection from 40% to 80%
- Implemented a tfidf and SVM (support vector machine) model for email classification and visualized the confidence scores distribution for unseen categories which beats the production model in terms of overall accuracy and robustness.
- Designed a statistical brand proximity metric which evaluate the product's user engagement and efficiency of the system response; A nonprovisional patent being applied in progress

09.2018-12.2018 **Machine Learning Engineer Intern,** Pactera OneConnect AI Lab New York

- Increased the slot-fitting rate of a task-oriented chatbot by adding a true-caser model in the NLP pipeline
- Implemented a naïve agender-based user simulator to provide large simulation data for building a reinforcement learning powered chatbot

06.2018-09.2018 **Machine Learning Engineer Intern,** Adatos A.I., New York

- Processed, augmented, and pipelined large batches of images to feed into the end-to-end Deep Unet model
- Decreased the mean absolute error (in around 300 trees) from 8 to 3 by replacing existing Gaussian blob detection with deep-learning powered tree counting model and speeded up the tree counting process from 1.83s per image to 0.1s per image by designing and fine-tuning an Unet Model
- Rescaled, trained, and transformed batches of images to solve the poor performance boundary problem of segmentation models (FCN and Unet)

EDUCATION

Columbia University *M.A. in Statistics, New York* GPA: 3.91/4.00 Sept.2017 - Jan.2019

Courses: Reinforcement Learning, Machine Learning, Analysis of Algorithm, Bayesian Statistics, Advanced Data Analysis

Fudan University *B.A. in Economics, Shanghai* GPA: 3.53/4.00 (Top 1 final year) Sept.2012 - July.2017

Courses: Applied Statistical Tools, Econometrics, Advanced Mathematics, Data Structures, Intro to Database

SKILLS

Model:	Lasso, Ridge, Logistic, Boosting, Bagging, CART, SVM, Clustering, Latent Dirichlet Allocation, EM algorithm
Tools:	Python, R, Bash, Docker, MongoDb, Numpy, SQL, Git, Scikit-learn, Keras, Pytorch, Tableau

PROJECT AND COMPETITION EXPERIENCE

04.2018 **Independent Project,** Collaborative Filtering with EM Clustering New York

- Built from scratches a recommender system with Bayesian Clustering Algorithm without using any ML framework
- Tried multiple parametrized distribution within the conjugate family, evaluated different models based on performance
- Use cluster structure robustness, perplexity to tune the hyperparameter and use utility score to evaluate the performance

04.2018 **Group Project Leader** Tweets Sentiment Analysis on Sino-US Trade War New York

- Implemented Latent Dirichlet Allocation to perform topic modeling for millions of tweets related to Trade War
- Use interactive D3 to visualize the retweet relation networks and found those influential opinion leaders on this topic
- Visualize majority opinion's shift across a series of events after performing sentiment analysis on all the comments

INTEREST

- Swimming (2000m nonstop with medley)