Tianyi (Mavis) Zhang

t.zhang116@rutgers.edu | (+1)17736068891 | maviszty.github.io

EDUCATION

Fordham University **New York, NY**
Master of Science in Data Science *08/2018-05/2020*

Rutgers University **New Brunswick, NJ**
Bachelor of Arts in Statistics, minor in Business Administration *08/2014-05/2018*

PROFESSIONAL EXPERIENCE

Researcher | iFlytek AI Research Institute **Hefei, CHINA**

AI Audiobook Narration System Development ***02/2023-Present***
Independent Project, advised by Vice President of Text to Speech research team
Developed a system for audiobooks that identifies character attributes to align with pronunciation databases

- Fine-tuned MacBERT for NER using EasyNLP framework, achieving a 99% F1 score in extracting character names and aliases from novels.
- Through Chinese RoBERTa, built a similarity model using PyTorch framework, and clustered character aliases by hierarchical clustering.
- Customized Chinese RoBERTa with full word masks and fine-tuned the customized model using the Pytorch framework to identify the character's gender and age, achieving 95% accuracy.
- Enhanced efficiency by switching from RoBERTa Large to MiniRBT and integrating models with Multi-Task, reducing reasoning time by 80%.
- Designed a prompt system for fine-tuning the iFlytek Spark Large Language Model(LLM) to recognize characters and analyze characters, extracting their attributes (gender, age) and deeper traits (temperament) automatically.

Data Science Lead | BigOne Lab Inc. **Beijing, CHINA**
As the first Data Scientist at BigOne Lab, built a data science team with 6 machine learning engineers for the company

NLP Public Opinion Analysis ***05/2021-12.2022***
Independent Project, advised by Dr.Wu (Former Google Brain Researcher)

- Developed a BERT NER model utilizing the Python TensorFlow framework with an f1-score of 0.91, identifying e-commerce trends from platform texts, resulting in our first consumer goods client, 50+ business expansions, and ~$2M annual revenue.
- Led creation of an ALBERT+TextCNN model, leveraging the Python-based PyTorch framework, categorizing ~100M e-commerce products, boosting data product accuracy from 50% to 95%, and saving 10,000+ annual work hours.

OCR E-commerce Image Recognition ***05/2021-12.2022***
Team Lead

- Created a Paddle OCR-based price recognition algorithm, tuned 20+ times for 90% accuracy in product price extraction; pioneered and deployed a deep learning price recognition system for top global consumer goods firms.

NLP Sentiment Analysis ***5/2021-12.2022***
Team Lead

- Researched and fine-tuned 10+ ML and deep learning sentiment models, achieving f1 scores over 0.9; delivered to 10+ clients within a year.
- Utilized PySpark and SQL to extract pertinent data from the database, employed KeyBERT for keyword extraction from Xiaohongshu, and conducted trend and sentiment analysis.

Media Mix Modeling ***12/2020-05/2021***
Independent Project, advised by Dr.Wu (Former Google Brain Researcher)

- Developed an XGBoost model utilizing the Python Scikit-Learn package, using advertisement cost to forecast daily buyers for leading cosmetic firms, achieving 99% accuracy during China's 2021 shopping festival.
- Designed an algorithm to enhance channel allocation efficiency, boosting cosmetics firms' monthly purchasers by 10% during the festival.

Data Scientist | Anduril Partners Inc. **New York, NY**
Factor Research ***2020/05 – 2020/10***
Independent Project

- Applied Random Forest feature analysis method on job posting and web traffic data to identify stock price determinants, offering actionable insights and recommendations to clients.
- Set up an AWS Sagemaker and EC2-based ML platform, trained UC Berkeley students on its use, and achieved 80% accuracy in predicting US equities revenue and 95% in stock price trend prediction.

Research Assistant | Fordham University Educational Machine Learning Lab **New York, NY**
Educational Data Mining Project (paper was published at the 2021 EDM conference) ***06/2019-05/2020***
Independent Project, advised by Dr.Weiss and Dr.Leeds

- Utilized Tmux to apply analysis on over 400,000 records of students' grades.
- Constructed Python library to automate analysis of correlations between courses and analysis of instructor grading styles
- Visualized course network by the directed graph in Gephi

Data Science Internship | BattleFin Group. Inc **New York, NY**
Predictive Modeling ***09/2019-12/2019***
Independent Project

- Performed data quality and integrity checks for alternative datasets before onboarding onto BattleFin's alternative data platform.
- Combined financial data with alternative data and fed it into deep learning models (GRU, RNN, LSTM) to predict stock price.

PUBLICATIONS

- **Educational Data Mining**: Daniel D. Leeds, Tianyi Zhang and Gary M. Weiss, Mining Course Groupings using Academic Performance. *Proceedings of The 14th International Conference on Educational Data Mining (EDM21)*, International Educational Data Mining Society, Paris France, June 29-July 2, 804-808.

PATENTS

- **Two Stage Character Attributes Identification**, Inventor: Tianyi Zhang, China Patent (Pending), Filed 11/25/2023, Developed and designed a system for identifying complex character attributes, aiming to improve the classification model's accuracy through denoising.

AWARDS

- **National Collegiate Taekwondo Competition**: Top 5 at the 2017 National Collegiate Taekwondo Competition; Represented Rutgers Taekwondo Club to take part in the training at West Point Military Academy
- **Deloitte Match Data Crunch Madness Competition:** 4th place out of 100+ teams

SKILLS AND LANGUAGE

- **Skills**: Python (*Pytorch, Tensorflow, Scikit-Learn*), SQL, R, Bash, JavaScript, Spark, Hadoop, Latex, Gephi, Tableau, SAS
- **Language**: Mandarin (Native), English (Fluent)

HOBBIES

- **Taekwondo**: Over 7 years of experience
- **Sketching**: Over 15 years of experience
- **Piano:** Level 8 Certification
- **Keyboard:** Level 10 Certification