Tian Yun

1121 Crowne Oaks Circle, Winston-Salem, NC 27106 | yunt16@wfu.edu | +1-530-551-9835

EDUCATION

Wake Forest University Winston-Salem, NC

Bachelor of Science in Mathematical Statistics *GPA:3.92/4.00* *August 2016 - May 2020*

Bachelor of Science in Computer Science *GPA:3.96/4.00*

- **Computer Science Relevant Coursework**: Database Management System, Parallel Computing, Machine Learning
- **Mathematical Statistics Relevant Coursework**: Statistical Learning, Time Series Analysis, Linear Algebra, Computational & Non-parametric Statistics, Ordinary Differential Equation, Categorical Data & Multilevel Models, Multivariate Statistics, Statistical Inference
- **Honor**: Dean's list all semesters; Member of Pi Mu Epsilon Honor Society; Member of Upsilon Pi Epsilon Honor Society; Member of Phi Beta Kappa Honor Society

RESEARCH EXPERIENCE

Natural Language Processing and Sentiment Analysis, Statistics Senior Thesis Winston-Salem, NC

Wake Forest University *August 2019 – Present*

- Conducted sentiment analysis on literature works and Twitter comments dataset using ideas of bag-of-words, n-grams, word hashing, TF-IDF, etc.
- Studied and derived latent Dirichlet allocation algorithm (LDA) and used R to finish coding my own version of LDA with Gibbs sampling process to label documents with latent topics
- Studying infinite latent variable model and aim at implement this to my LDA function to find the optimal number of latent topics
- Thesis: *A Brief Introduction to Natural Language Processing*
- Link to thesis: https://github.com/tttyuntian/Statistics-Thesis

Object Classification and Segmentation, Computer Science Honor Program Winston-Salem, NC

Wake Forest University *September 2019 – Present*

- Studied methods for image data preprocessing and data augmentation
- Studied convolutional neural network and used Keras to constructed U-Net to segment the map-related objects (e.g. roads, building, tracks) in satellite maps
- Set up experiments to explore the relationships between the parameters (i.e. learning rate, batch size, etc.) and the performance of U-Net in segmentation and detection tasks

WORK EXPERIENCE

Tencent Holdings Limited Shenzhen, China

Data Analyst Intern *May – August 2019*

- Extracted relevant data from HDFS (over 10 billion rows of data per day) and processed data to monitor the performances of QQ Kandian and Kandian APP using Hive-SQL or Spark-SQL
- Used Flask, Python and Git Webhook to develop back-end service to listen POST requests from the created Git repository to help colleagues to upload the scripts to the task scheduling system (i.e. help to create tasks and to set up task dependence) efficiently
- Evaluated the causal relationships between the loss of daily active users and the potential factors; predicted if the daily active users will leave in the following week by fitting models, like the random forest, gradient boosting decision tree (GBDT), etc., by using PySpark and Python (Currently, the best model has AUC score of 0.9173)

China Minsheng Bank Beijing, China
Data Analyst Intern *May – July 2018*

- Aggregated and evaluated data for 8 million customers and obtained 500 thousand records with all the information of the customers (i.e. over 200 features) using SQL on DBM2
- Fitted GBDT, random forest, logistic regression models in Spyder to find the target customers (i.e. 48,000/8 million) for the bank to sell the financial products

TUTORING EXPERIENCE

Computer Science Peer Tutor, Wake Forest University Winston-Salem, NC
Tutor *September 2017 – Present*

- Provided tutoring sessions from low-level to high-level computer science courses to explain the class materials in detail or to provide assistance on assignments and test reviews

AWARDS

ICPC Mid-Atlantic USA Regional Contest 2019, UNC – Chapel Hill Chapel Hill, NC
4th Place *November 2019*

- Designed and implemented algorithms in Python to solve problems

ASA DataFest 2019, Duke University Durham, NC
Honorable Mention *April 2019*

- Created Fatigue Composite Index (FCI) to quantify the individual variation in subjective self-reported data from athletes in Canadian National Women's Rugby Team
- Used FCI to fit bagging, random forest, lasso regression models to predict measures of wellness and evaluate the importance of predictors
- Discovered that the fatigue level of an athlete is influenced by the type of training session he/she attends

COMAP's Mathematical Contest in Modeling 2019 Winston-Salem, NC
Honorable Mention *January 2019*

- Collected environmental, demographic, and building-area-related data from 217 countries from a variety of websites
- Fitted random forest and regression models to understand the relationships between the created Environmental Degradation Index and the predictors and to quantify the environmental degradation costs

SOFTWARE AND LANGUAGE SKILLS

Programming Language: Python, Java, R, C++, C, SQL, Hive SQL, Spark SQL, PostgreSQL, Matlab
Software: PyCharm, Spyder, Jupyter Notebook, RStudio, Tableau, DBM2, SQLiteStudio, Microsoft Excel, Word, PowerPoint
Python Library: scikit-learn, numpy, pandas, keras
Language: Fluent in English, Mandarin, and Cantonese