WENKANG WEI

Current Address: Central, South Carolina, 29630 || **Phone**: 864-324-4885 || **Email (Preferred)**: wenkanw@g.clemson.edu
Tech Blog: wenkangwei.github.io **LinkedIn**: www.linkedin.com/in/wenkang-wei-588811167/ || **Academic Profile**: meritpages.com/wenkan

SUMMARY

I'm a second-year master student with 3 years of experience in Machine Learning and 5 years of experience in programming, actively looking for a full-time job of **Machine Learning Engineer / Researcher**, starting from Summer 2021.

TECHNICAL SKILLS

Programming: Python / Jupyter Notebook, PostgreSQL, C/C++ , Matlab, HTML, Markdown, Latex
Tools: PyTorch, Tensorflow, PySpark, MPI, Hadoop MapReduce, Data Analysis toolkits (sklearn, pandas, seaborn, etc), Raspberry Pi, Linux, Google Colab, Git

Theory and Analysis Techniques:

- Feature Engineering, Data Visualization and Preprocessing Techniques, PCA, Word Embedding, TF-IDF, etc
- Machine Learning Modeling: Collaborate Filtering, Matrix Factorization, SVM, Decision Tree, Clustering, Convolution Neural Network etc
- Techniques for Model Evaluation and Improvement: Cross-Validation, Ensemble Learning, ROC, AUC, Feature Importance, etc.

WORK EXPERIENCE

Machine Learning Research Assistant Clemson University Summer 2020-Current

- **Proof of convergence and convergence rate of Multiple Update Algorithm (MUA) in Non-Negative Matrix Factorization Problem**
  1. Formulated Matrix Factorization Problem into Constraint Optimization Problem
  2. Applied Linear Algebra, Lagrange multiplier to simplify problem and utilized Lipschitz gradient, convex optimization to prove the convergence and convergence rate of MUA algorithm
  3. Implemented MUA and ALS (alternative least square) algorithm in Google Colab and Matlab to verify convergence result
  4. Collaborated and communicated with CS professor to present mathematic proof process orally
  5. Wrote a paper in AAAI format using Latex (unpublished due to copyright)

EDUCATION

Master of Computer Engineering,	Clemson University, Clemson, SC	May 2021	*GPA 3.71/4.0*
Minor: Computer Science	Clemson University, Clemson, SC	May 2021	
Bachelor of Electrical Engineering,	Clemson University, Clemson, SC	May 2019	*GPA 3.8/4.0*

Coursework: *Applied Data Science, Data Mining, Distributed Computing, Reinforcement Learning, Nonlinear control system, etc.*

SELECTED PROJECT EXPERIENCE

Car Classification using Transfer Learning Fall 2020

- Constructed data pipeline by **PyTorch** to extracted and transformed Stanford car images dataset (1.96GB dataset with 196 classes)
- Modified and tuned pre-trained models Google-Net, VGG-16 , Res-Net 50 to fit car dataset using early stopping, weight decay techniques
- Improved test accuracy of the best model to **85%** using cross-validation model selection techniques

Recommendation System based on MovieLens 25M dataset Fall 2020

- Utilized **PySpark** to load movielens 25M dataset (25 million ratings) and used SQL to query and analyze data in databrick cluster platform
- Implemented and applied Mapper, Reducer functions in **Hadoop** File system to analyze contribution of different movie genres to ratings
- Applied Collaborative Filtering and Matrix Factorization methods to construct a recommendation system with PySpark
- Achieved 0.67 mean square error score and deployed recommendation system using IPython widget

Youtube Comments Analysis and Pet Owners Classification Fall 2020

- Utilized PySpark and SQL to load, query and explore Youtube comment text data (about 1GB after decompression)
- Built data pipeline and applied Term-Frequency-Inverse Document-Frequency(TF-IDF) to transform text data into numerical data
- Applied Logistic Regression, Random Forest, Gradient Boosting machine in PySpark to classify cat or dog owners from comments.
- Achieved **92%** prediction accuracy on test set using Grid Search and Cross Validation method

Improvement on Bank Customer Churn Prediction Summer 2020

- Visualized and analyzed data related to customer churn by using visualization toolkits: seaborn, matplotlib
- Preprocessed and transforms categorical data for Machine Learning model training using pandas toolkit and normalization techniques
- Established Data Pipeline and ML Models: Random Forest, Logistic Regression, SVM, etc. and Evaluated Models using ROC,AUC
- Improved Models Accuracy from **80% to 86%** by using Model Selection, Cross Validation and Feature Selection, L1 Regularization techniques

IMDB Movie Rating Classification Summer 2020

- Collected text data using BeautifulSoup tool and cleaned data by Stemming, removing stop words to analyze 1.4GB movie rating dataset
- Applied **Word Embedding**, **Bag of Word** model**, TF-IDF** Techniques to transform text data into different representations for model training
- Designed Convolution Neural Network with **Tensorflow** and applied other models: Support Vector Classifier, Random forest, etc.
- Evaluated model performance on validation, test dataset and achieved model testing accuracy **88%** by model tuning

Quora Sincere Question Classification Kaggle competition Spring 2020

- Communicated and collaborated with teammates to allocate work and balanced workload
- Analyzed and visualized frequent keywords in **6GB Quora dataset** using world cloud to identify keywords related to insincere/sincere question
- Cleaned text data by missing values filling, stemming, stop words removing and visualized word clusters after applying Principle Component Analysis (PCA) to reduce data dimension
- Transformed text data into word-2-vector, TF-IDF and constructed models: BiLSTM, Random Forest, SVM, etc to classify sincere questions

Smart Reminder in CU-HackIt competition Spring 2018

- Proposed ″Smart Reminder“ idea to improve worker productivity real world problem and collaborated with teammates to balanced workload
- Assisted to constructed Flask IoT Server in Raspberry Pi and designed Servo Motor, Buzzer control algorithm to interact with users
- Applied OpenCV for image processing and Tensorflow pre-trained model on COCO dataset for human detection
- Achieved the ***Best overall Hack Price*** with confident teamwork

HONOR

Eta Kappa Nu (HKN) (Spring 2018 - present); Golden Key Honor Society (Fall 2017- present); President’s List (Summer 2019);

Dean’s List (Fall 2016 – May 2019); Best overall hack award of CUhackit Competition (Spring 2018);