Luwei (Caroline) Lei

leiluwei4221@gmail.com | (814) 862-8939 | https://www.linkedin.com/in/caroline-luwei-lei/

EDUCATION

Georgetown University Expected Graduation: May 2021
M.S. in Data Science & Analytics **GPA: 4.00/4.00**, Merit-Based Scholarship
Coursework: Data Structure and Algorithms, Statistical Learning, Cloud Computing, Visualizations, Deep Learning

The Pennsylvania State University Aug 2015 - May 2019
B.S. in Computational Data Science, *B.S. in Mathematics* GPA: 3.61/4.00, Dean's List
Course: Machine Learning, Computer Vision, Artificial Intelligence, Database Management, Linear Algebra

SKILLS

Tools: Python (Sklearn, NumPy, Pandas, Keras, PySpark, Tensorflow), R (dplyr), SQL (MySQL, BigQuery), Visualization (Tableau, Matplotlib, Seaborn, Plotly, Shiny), Hadoop, Spark, HTML, Excel, Google Analytics
Skills: Machine Learning, Deep Learning, NLP, Data Warehousing, A/B Testing, Statistical Test, Experimental Design
Interests: Cooking, Theatre art, Painting (volunteer at local school), Fashion, Hiking

PROFESSIONAL EXPERIENCE

Data Scientist Intern, *Center for Security and Emerging Technology, Washington, DC* June 2020 - Present

- Constructed automated Python scripts to retrieve data from BigQuery, identified data issues (inc. anomalies, mistaken values) and created reports and visualizations for multiple large real-world databases to validate the data correctness for research team
- Designed and developed Apache Beam pipelines to normalize data issues, regularize data formats and perform specific ETL tasks (inc. query extraction) on more than 1 TB high dimensional databases for research team
- Performed exploratory data analysis to identify patterns in PhD student stay rates in U.S., highlighted changes in top countries, visualized findings using Plotly and presented summarized insights to scientists
- Contributed to a NLP research on automated study of language surrounding AI by parsing Chinese corpus, analyzing language features (inc. mutual information, co-occurrence words) and submitted a paper to ACL conference

Software Engineer Intern, *SenseTime*, *Shanghai, China* May 2019 - Aug 2019

- Led the team to design and implement a Python educational game and successfully promoted to 20+ middle schools in China
- Actively communicated the front-end and external personnel to advance the game platform to get published
- Tested and identified optimization opportunities of voice recognition and multi-armed bandit game performances via Pytorch
- Participated in an algorithm textbook editing project and wrote sorting algorithm, which was used by 2000+ students
- Constructed image classification algorithm using machine learning models in Pytorch with 90%+ accuracy and created user friendly API to 20+ educational organizations

Data Research Assistant, *The Pennsylvania State University, State College, PA* May 2018 - Aug 2018

- Performed ETL on U.S. government petition dataset and tokenized over 7,000 petitions into sentences using NLTK package
- Extracted and integrated natural language processing (NLP) features of the sentences including word embedding, dependency parsing, LSTM and POS tag via Python NLTK package
- Implemented a convolutional neural networks (CNN) to classify petitions based on semantic meanings and structures, and improved the model performance by 10%

SELECTED PROJECTS

Real-time Face Mask Detection, *Georgetown University* Aug 2020 - Dec 2020

- Composed and trained multi-layer CNN, MobileNetV2 and VGG16 models using Keras to classify headshots and compared performance of all the models, the best model achieved 92% accuracy
- Built a real-time face recognition application in OpenCV and embedded the best trained model

Third Prize of COVID-19 Data Challenge, *CGDV by QED Group* May 2020

- Identified the impact of COVID on different social aspect and gathered relevant data from variety of sources
- Constructed different interactive visualizations using Dash and Plotly to explore the impact of COVID-19 and government responses on mobility, public opinion, unemployment and legislation in U.S., built a website to present overall findings

Social and Economic Impacts of Broadway Shows, *Georgetown University* Aug 2019 - Dec 2019

- Scraped over 50,000 records of weekly gross data and textual reviews of Broadway shows via Python, performed exploratory analysis and statistical testing to identify the impacts of seasonality, holidays and social media on show sales
- Developed topic modeling analysis via Python Gensim package to unveil the critics' preferences of shows

Improving Rugby Game Performance, *Winner of 2019 ASA DataFest at Penn State University* April 2019

- Translated customers' needs into high-level data models, designed metrics to evaluate game performance of the team, standardized and normalized biased features, and trained a random forest model to measure the important features of the game
- Developed actionable strategies based on findings of the project and provided the personalized suggestions to each player