# Kuan-Lin Liu

kuanlin.liu@nyu.edu ◇ (929) 332-5669 ◇ Brooklyn, NY ◇ linkedin.com/in/kuanlinliu

## PROFESSIONAL EXPERIENCE

**Bombora** New York, NY
*Data Scientist Intern* Sept 2020 - May 2021

- Set up an end-to-end active learning pipeline on 1M web contents with BigQuery for extracting more balanced samples
- Saved 70% labelling costs by implementing uncertainty-based querying methods based on **fastText**'s probability score
- Launched the Embedded Topic Model on **Google Cloud Platform** to capture topics and identify B2B contents
- Detected unexpected data by pre-processing and visualization (**Pandas**, **Seaborn**) to revise the web-scraping process
- Generated tree-based decision rules with topic modelling to speed up human's judgement on data labelling by 60%

**New York University** New York, NY
*Graduate Research Assistant* May 2020 - Mar 2021

- Designed Seq2Seq attention networks to adaptively model printed texts into phonemes according to images' difficulty
- Performed distributed computing with CUDA and multiple GPUs through PyTorch Lightning to accelerate training
- Examined model workflow with unit testing (**Pytest**) and visualized computation time in FLOPs on each pixel

**Zillow Group, Inc.** New York, NY
*Master's Capstone Member* Sept 2020 - Dec 2020

- Enhanced average recall by 47% compared to Amazon Textract for predicting key-value pairs on 2200 utility bills
- Boosted document understanding models (**LayoutLM**, **BERT**) with weighted loss to handle high class imbalance
- Developed a prototype web app (**FastAPI**, **Streamlit**, **Docker**) and organized teamwork via **Git** version control

**Cathay Financial Holdings Co., Ltd.** Taipei, Taiwan
*Data Science Intern* July 2020 - Aug 2020

- Improved model's specificity by 35% and TextBlob's biased target distribution by **Graph Convolution Networks**
- Built an ETL pipeline along with Python Multiprocessing to accelerate text cleaning on 700K E-commerce reviews

## TECHNICAL SKILLS

| | |
|---|---|
| **Programming** | Python, Scala, SQL, R, Bash Script, C++, HTML, CSS |
| **Software & Tools** | Hadoop, Spark, AWS, GCP, CUDA, MySQL, Git, Docker, RESTful API, Tableau |
| **Frameworks & Packages** | PyTorch, Scikit-learn, Pytest, FastAPI, Streamlit, Flask, Dash, Matplotlib |

## EDUCATION

**New York University** New York, NY
*M.S. in Data Science, GPA: 3.79 / 4.0* Sept 2019 - May 2021

- *Selected Courses*: Deep Learning, Natural Language Understanding, Big Data (Hadoop, Spark, Scala), Machine Learning, Tools and Techniques for Machine Learning, Database Systems, Introduction to Data Science

**National Taipei University** New Taipei, Taiwan
*B.B.A. in Statistics and B.A. in Economics, GPA: 3.83 / 4.0* Sept 2014 - Jan 2019

- *Selected Courses*: Algorithms, Data Structures, Data Mining, Dimension Reduction, Regression Analysis, Time Series

## SELECTED PROJECTS https://garylkl.github.io/#portfolio

**Book Recommendation System for Goodreads.com**

- Analyzed 4GB user-item interaction data in **PySpark** and built a Collaborative Filtering recommendation system
- Increased efficiency of the book-searching system by 10 times and visualized clustering of products with T-SNE

**Fake Reviews Detection for Restaurants**

- Led the text preprocessing pipeline to boost efficiency from 20 hours to minutes by **Spark NLP** and **NTLK**
- Engineered on behavioral features to improve ROC score by 16% (XGBoost) against text features (**GloVe**, **BOW**)

**Kickstarter's Success Prediction and Product Recommendation**

- Developed an ETL pipeline using **Spark SQL** and **Scala** on HDFS and Spark for faster cleaning JSON files
- Built Machine Learning pipelines with **Spark's MLlib**, and visualized patterns with a Tableau dashboard
- Recommended similar products on Amazon by Locality-sensitive Hashing and extracted top words by TF-IDF