Tim Hsu

Phone: (217) 402-4762 Email: juiting.hsu@nyu.edu

EDUCATION

New York University **New York, NY**
May 2018 3.8/4.0
Master of Science, **Data Science**

University of Illinois at Urbana-Champaign **Urbana, IL**
Bachelor of Science with Honors, **Computer Engineering** 3.7/4.0
Bachelor of Science with Highest Distinction, **Statistics**

Courses:

Data Structures
Database Systems
Machine Learning
Linear Algebra
Natural Language Processing
Algorithms
Deep Learning
Big Data

SKILLS

Language	Python, Java, SQL, R, Objective-C, SQL
Library/Framework	Scikit-Learn, NLTK, Torch, Flask, Tensorflow, Spark, etc
Systems/Platform	Linux, AWS, Tableau, MapReduce, Hadoop, Git

JOB EXPERIENCE

BlackRock **New York, NY**
Data Scientist Intern **Summer 2017**

- Proposed and implemented a solution for legal clause identification in contract documents by using word embedding trained with Word2Vec model and unsupervised learning algorithms to distinguish text blocks associated with certain legal provisions from one another. Labels are then inferred by inspecting the group and applying domain knowledge.
- Designed and implemented a solution a RESTful API using Python's Flask framework that includes retrieving and updating info about contracts in database and repository, generating on-the-fly result using the trained model, and providing a feedback system for users to indicate correctness, etc.

PROJECTS

Quora Paraphrase Detection [Pytorch, Python] **Fall 2017**

- Implemented a Bilateral Multi-Perspective Matching neural network using PyTorch (Python) and pretrained GloVe word embedding to perform paraphrase detection task. The model takes in two sentences (or phrases) as inputs and predicts whether they are paraphrases of each other or not.
- Evaluated result and compared model with different settings on the Quora question pairs dataset. Tuned hyperparameteres on the validation set and achieved a test accuracy of 88%.

NYC Crime Data Analysis [Hadoop, Spark, AWS] **Spring 2017**

- Preprocessed and cleaned NYPD Crime dataset from 2015 containing 5.58 million rows and 24 columns of data recording crimes reported to the NYPD from 2006-2015 using Hadoop MapReduce and PySpark.
- Incorporated relevant dataset such as restaurants and weather to develop new metrics to investigate relationship between crime and area. Showcased result using data visualizations such as box plots, bar charts, and geo-location maps from Python libraries and Tableau, compiled with a presentation to tell the stories.

Movie Recommendation System [Scikit-learn, pandas] **Spring 2017**

- Implemented recommendation system algorithms such as matrix factorization with ALS and SGD, and content-based filtering on the Movie Lens dataset in order to recommend movies by predicting users' ratings from 1 to 5.
- Preprocessed dataset to provide an evaluation scheme for the team by splitting the data into training, validation, and test set while considering their timestamps to model a realistic setting. Achieved RMSE of around 0.91, compared to Netflix's 0.88.

Handwritten Digit Recognizer [Pytorch, CNN] **Summer 2016**

- Trained a CNN model to perform handwritten digit classification using PyTorch in a supervised fashion on a dataset with only 3000 labeled data and 47000 unlabeled data as a baseline model to achieve a test accuracy of 95%.
- Used semi-supervised methods such as pseudo-labeling, data augmentation, and implementing a ladder network to improve the test accuracy from 95% to 98.3%.