YUJI SHEN

DATA DEVELOPER

yuji.im.shen@gmail.com (412) 228–1560 http://yuji.im https://github.com/YujiShen

EDUCATION

University of Pittsburgh, *M.S. Information Science* 2015

GPA: 3.8. (Database Management, Information Retrieval, Data Analytics, Information Visualization, Data Mining, etc.)

Springboard Data Science Workshop, *Data Science Mentee* 2015

Mentored by industry experts to learn the foundation of data science and built portfolio site with real-world projects.

Coursera, EdX, Udacity, *Data Science MOOCs* 2015

Probability, Statistical Inference, Analytics Edge, Machine Learning, Apache Spark Big Data Series, etc.

Hunan University, *B.B.A. E-Commerce* 2013

SKILLS

LANGUAGES: Python, Java, SQL, R, HTML, CSS, JavaScript, Git, Regex
OTHERS: Apache Spark, Pandas, Matplotlib, ggplot2, D3.js, AWS, Linux

PROJECTS

Personal Time Analysis Pipeline, *Personal Project* 2015

- A personal data pipeline to get data, visualize statistics and generate reports, using Python, MySQL.
- Retrieved recent time records from aTimeLogger API and updated MySQL database.
- Aggregated and analyzed data, plotted in graphs and tables in Python (Pandas, Matplotlib).
- Created structured note with plot in Evernote Python API and insert into my account.
- Automated personal review process, got 5 starts on GitHub, including the developer of aTimeLogger.

Display Advertising Prediction, *Personal Project* 2015

- A Kaggle data science competition to predict click through rate, using Apache Spark, Python, AWS.
- Deployed Spark cluster on AWS Linux (CentOS) machines, built a Network File System for sharing data within cluster.
- Decreased feature sparsity over 300 times with preprocessing and Hash Algorithm.
- Accelerated Logistic Regression with Limited-memory BFGS optimization methods.
- Improved model with Grid Search, beat benchmark and ranked top 30% on leaderboard.

Pittsburgh Neighborhoods Analysis & Visualization, *Team Project* 2015

- A statistical report and visualization site to represent the analysis of Pittsburgh neighborhoods, using R, D3.js.
- Utilized statistical method to analyzed demographic and economic data of neighborhoods in Pittsburgh.
- Learnt D3.js in 3 weeks and implemented Vizburgh website, used HTML5 Local Storage for multiple profiles data accessing.

Twitter Hashtag Search Engine, *Team Project* 2014

- A search engine to recommend Hashtag according to user's input, using Java, Apache Lucene.
- Coded in Java to extract and process fields for building inverted index, based on 50 million tweets.
- Utilized Apache Lucene with Vector Space Model and Boolean Model (default) for searching.
- Optimized result relevance by combining scores from different fields and using fuzzy query.

WORK EXPERIENCE

CloudParticle, *Developer Intern* 2015 - Current

- Optimized database design with normalization and queried data using SQL.
- Cleaned and processed data in Python, analyzed and visualized result.

Changsha Excellent Internet Info Service, *Student Intern* 2013

- Led a group of interns to collect, clean and select data for marketing section, using Java and batch script.