Di Chen

(984)-269-8298 | http://dichen.me | dchen20@ncsu.edu
905 W Middlefield Rd #976, Mountain View, CA 94043

EDUCATION

North Carolina State University (NCSU), Raleigh, NC — Aug. 2015 - May. 2018
M.S. Computer Science | Supervisor: Dr. Timothy J. Menzies, Full professor

University of Science and Technology of China (USTC), Hefei, China — Sep. 2011 - Jun. 2015
B.S. Electronic Engineering and Information Science | Supervisor: Dr. Zuqing Zhu, Full Professor

SKILLS AND INTERESTS

- Experience in Machine Learning, Natural Language Processing, Graph Mining, Distributed Systems and DevOps;
- Proficient in *Python*, *C/C++*, familiar with *Java, JavaScript, R*, good at most common development and data analysis tools;
- Interested in backend/infrastructure development as well as machine learning, text mining, and research positions.

INTERNSHIP EXPERIENCES

Software Engineer Intern | **Google, Mountain View, CA**
Large Scale Graph Mining on User Behaviors for User Intents Discovery — May. 2017 - Aug. 2017

- Working on Google AdWords. The project aimed to 1) understand user intent from a large amount of aggregated user behaviors associated with queries 2) find queries that are highly related to certain targeted activities.
- Contributions include 1) Building and mining user behavior data with graph mining techniques(e.g. label propagation) to find clusters of similar behavior patterns and the associated queries. 2) Applying label propagation on user behavior graph to improve our current models. 3) Building an active learning pipeline to reduce the time and costs for annotating new training data. 4) Implementing a Flume pipeline in *C++* to extract alarge amount of internal data in parallel.

Software Engineer Intern | **LexisNexis, Raleigh, NC**
Developing Large-Scale User Profile Deduplication Algorithm — Sep. 2016 - Dec. 2016

- Aimed to deliver a machine learning algorithm to effectively identify duplicate profile in large scale, with the constraint of limited training data and partial information. Focus on the both the efficiency and accuracy of the deduplication algorithm.
- Contributions include 1) Collecting training data from *SQL* servers storing the change history of user profiles. 2) Creating baseline results by evaluating the existing rule-based algorithm with *pandas*. 3) Extracting, building and selecting features to improve model's precision and reduce time complexity with *scikit-learn* and *Weka*. 4) Testing performance with data pulled from a different database and evaluating via crowdsourcing. 5) Data visualization with *D3.js* and *Plotly*.

Software Engineer Intern | **LexisNexis, Raleigh, NC**
Machine Learning for Email Signature Extractor and Profile Parser — May. 2016 - Aug. 2016

- Aimed to develop an ML based Email signature extractor for user profile building, so as to replace the Regular Expression based one in production. Focus on building models capable of transfer learning and improving precision and recall.
- Contributions include 1) Data collection via crowdsourcing for supervised learning; 2) Active learning for continuous model improvement; 3) Deep Learning with *word2vec* model trained on *Google News* for word embedding and feature engineering; 4) Model precision and recall improved 10% and 20% respectively, compared with the previous model.
- Code was reviewed, perfected, and launched to production. Models were introduced to be used in multiple projects.

SELECTED PROJECTS

Research Assistant | **RAISE Lab, NC State University**
Mining and Understanding GitHub Pull Requests — May. 2016 - Jan. 2017

- Implemented a dynamic crawler with *Beautiful Soup* to scrape *GitHub* in parallel for data inaccessible via *GitHub API*.
- Applied qualitative approach for feature extraction and followed by quantitative analysis for model building.
- Built a pull-request fate predictor achieving above 90% precision and recall. Also built a reviewer recommendation system achieving 80% precision for top 3 recommendation. Learned insights will contribute to improving software development on open source community. This work has been written into paper and submitted to TOSEM.

Innovation Design Project | **Microsoft Research Asia & USTC**
Web Application for Easy Communication — Sep. 2014 - Jun. 2015

- Designed and implemented a video chat web application with voice control and facial recognition for the old in China.
- Mainly contributed to applying and integrating speech recognition, facial recognition, and page auto generation altogether in *Java*. Also involved in developing the applications interface and user interaction with *JQury, Bootstrap*, and *MongoDB*.

Research Assistant | **Networking System & Internet Infrastructure Lab, USTC**
Algorithm Design and Simulation for Software Defined Optical Networks — Jun. 2014 - Sep. 2014

- Designed the protection schemes for multipath provisioning(*MPP*) to optimize the bandwidth efficient while guaranteeing 100% restoration against single-link failures in elastic optical networks(*EONs*) with the Maximum Independent Set.
- Implemented simulations on *Matlab* and conducted a performance evaluation. Results are published to *GLOBECOM*.