Boston, MA Email: sanwang@gwu.edu

San (Samantha) Wang

GitHub blog: https://san-wang.github.io/ • LinkedIn: linkedin.com/in/san-wang

EDUCATION

The George Washington University (GWU) DC, 01/2016-01/2018

M.S. in Data Science, GPA: 3.9/4.0

- Featured Classes: Machine Learning I & II, Data Analysis, Data Mining, Data Warehousing, Visualization of Complex Data

Sichuan University Chengdu, China, 09/2011-06/2015

B.S. in Mathematics, concentration in Statistics, GPA: 3.2/4.0

- Outstanding Undergraduate Thesis (5%); Second Class Scholarship in 2014

SKILLS

- Software: Python (Keras, scikit-learn, Tensorflow, Caffe, EDA), Spark, Tableau, postgresql, MySQL, neo4j, GCP, Git, Linux
- Technique: ETL, Computer Vision, NLP, Recommendation System, Machine Learning Modeling, Data Mining, Data Visualization

WORK EXPERIENCE

Machine Learning Research Associate MA, 01/2019-06/2019

Harvard Medical School (HMS)

- Main contributor for developing image fraud screening system to detect improper reuse of medical images using Keras
- Utilize image feature extraction, metric learning, and convolutional neural network techniques; overcome the challenge of detecting dissimilar images with regional fraud by designing a novel Siamese neural network
- Distance-based model achieves 89% accuracy and novel non-distance based model receives 90% accuracy on real world test
- Cooperate with principal investigators to improve metadata collecting process and improve data storage structure
- Work was selected to present at Women in Data Science conference poster session in 2019

NLP Research Assistant (Blog) DC, 05/2017-01/2018

The George Washington University (GWU)

- Led 4 graduate students to analyze top technical skills demand, market demand and supply of data scientists across the U.S.
- Collected supporting data by scraping; applied stemming, synonym dictionary, stop-words removing and TF-IDF technique to optimize NLP approaches, analyzing technical skill requirement and opening availability for data-related positions
- Built interactive Tableau reports, such as heatmap, Sankey chart, word-cloud, geographic map, presenting findings efficiently and in a non-technical audience-friendly way, receiving 1.7k views
- GWU data science technical blog and DC DataCon 2017 conference reported our team's research to demonstrate the insights of the data science industry and the gap between academic curriculum design and industry requirements

PROJECTS

Fashion Product Discovery Platform(theGlancer) MA, 12/2018-03/2019

- Core data scientist member to empower Glancer app with the power of machine learning technique for tagging products, generating titles for mobile display, providing product recommendations using Python
- Develop product tagging pipeline with keyword matching system and multi-label classifier using deep neural network
- Extract key information from raw data collected from API and transform to bag of words representations using TF-IDF

Movie Recommendation System [Python/Spark/Flask] (Demo) MA, 03/2018-10/2018

- Built ETL pipeline to conduct online analytical processing (OLAP) with Spark SQL to analyze user and item profile
- Created customized content-based NLP model to provide recommendation based on movie plot and cast preference
- Conducted Collaborative Filtering (CF) models for personalized recommendation using ALS matrix factorization in Spark MLlib, provided user-based CF model to handle item cold start
- Designed and developed a Flask web API backed with Postgres database to provide user-friendly UI

Avito Product Demand Analysis [Python] MA, 05/2018-06/2018

- Predicted product demand for an online ads platform by analyzing users' behavior pattern and products' information
- Trained various supervised learning models, such as Random Forest, lightGBM, logistic regression, and applied regularization technique to overcome overfitting, identifying most important factors that affect product demand