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, 01/2016-12/2017

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

- Featured Classes: Data Analysis, Data Mining, Data Warehousing, Machine Learning I&II, 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 (Tensorflow, keras, scikit-learn, Caffe, EDA), Spark, Tableau, postgresql, MySQL, R, Google Cloud Platform, Git
- Technique: ETL, NLP, Recommendation System, Neural Network, Modeling, Data Mining, Data Visualization, Data Cleaning

WORK EXPERIENCE

NLP Research Assistant (Blog) DC, 05/2017-12/2017

The George Washington University (GWU)

- Led 4 graduate students to analyze top technical skills demand, market demand and supply of data scientists across U.S.
- Collected supporting data by scraping webpages using Python Beautiful Soup and Regular Expression
- Applied stemming, synonym dictionary, stop-words removing and TF-IDF technique to optimize NLP approaches to analyze technical skill requirement and opening availability for data scientists, data engineer and data analyst positions
- Built interactive Tableau reports, including but not limit in heatmap, Sankey chart, word-cloud, geographic map, to demonstrate findings efficiently and non-technical audience friendly
- GWU data science technical blog and DC DataCon 2017 conference reported our team's research to demonstrate the insight of data science industry and the gap between academic curriculum design and industry requirement

PROJECTS

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

- 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, similarity-based CF model; provided user-based CF model to handle item cold start
- Designed and developed a Flask web API backed with Postgres database to provide users with customized movie recommendations and movie exploration by genre and ratings

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

- Predicted product demand for an online ads platform by analyzing users and products' unstructured (title and description in Russian, images) and structured data (price, category, location and time)
- Transformed raw text to bag of words representations using TF-IDF; extract image bottleneck features from pre-trained VGG16 model; Imputed important features' missing value using regression
- Trained various supervised learning models, and applied regularization and dropout technique to overcome overfitting, improving RMSE score by 6% in test dataset
- Built pipelines to automate preprocessing and hyper-parameter tuning for Random Forest, lightGBM, logistic regression
- Evaluated model performance using k-fold cross-validation and identified most important factors that affect product demand

Master Capstone, Image Classification [Tensorflow/Caffe/Keras] (Blog) DC, 02/2017-12/2017

- Classified 39K traffic sign images in 43 categories using convolutional neural network(CNN) in Keras, Tensorflow and Caffe
- Analyzed CNN black box by visualizing kernels, feature maps and building customized animation plots to show the change of all kernels during training using Tensorboard
- Reduced training time using transfer learning and GPU in Google cloud platform
- Gained 90.58% accuracy when testing on 12K images in Caffe and 95.42% in Tensorflow and Keras