Boston, MA Email: sanwang@gwu.edu

San (Samantha) Wang

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

SKILLS

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

WORK EXPERIENCE

Computer Vision Research Associate MA, 01/2019-06/2019

Harvard Medical School (HMS)

- Main contributor for developing image manipulation detection system with quantitative assessment using machine learning
- Cooperated with principal investigators to improve metadata storage structure and relational database schema
- Developed features including model evaluation with hierarchy classification report and result visualization, Siamese neural network pair-generation policies with training difficulty level control, data checking pipelines to spot corner cases
- Overcame the challenge of detecting dissimilar images with regional fraud by digging into literature, accessing related work and designing a novel non-distance based Siamese neural network
- Distance-based model achieves 89% accuracy and novel non-distance based model receives 90% accuracy on real-world test
- Selected to present at 2019 Women in Data Science conference poster session

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 share across the U.S. using various job media sources
- Collected and structured supporting data through web scraping; applied stemming, synonym dictionary, stop-words removing and TF-IDF technique to optimize NLP approaches, analyzing trends and job market share in geographical distribution
- 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

EDUCATION

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

M.S. in Data Science

- 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

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

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 indexing products, generating titles for mobile display and exploring trending new products
- Collected raw data through Rakuten API; extracted key information from JSON; maintained updated daily with relational database and neo4j; prepared product profiles for analysis and training machine learning models
- Develop product tagging pipeline based on keyword matching and multi-label classifier models by analyzing products' image and text information, including description, title and brand name

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 behavior and trending pattern
- 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 and mySQL database to provide user-friendly UI