Du Xiang

+1 (858) 666-5552 | dxiang@ucsd.edu | Portfolio: http://xd00099.github.io

EDUCATION

University of California – San Diego, Halıcıoğlu Data Science Institute, California, CA 2018 - 2022
Bachelor's in Data Science | Minor in Management Science | **GPA4.0**

TECHNICAL STRENGTH

Programming Languages: Python, Java, SQL, Bash, Javascript, HTML/CSS, R, MATLAB
Software & Tools: Pandas, scikit-learn, Jupyter, Docker, Neo4j, PostgreSQL, Apache Spark, Dask, AWS, Matplotlib, Seaborn, Keras, Git, Tableau, Matlab, Jupyter, TigerGraph, Anaconda, Github, Vim
Text mining & NLP: Scrapy, Beautifulsoup, Word2Vec, Doc2Vec, BoWs, Tfidf, Gensim Topic Modeling (LDA), XML Parsing, Bert tokenizer, NLTK, N-Grams, Mallet

COURSES

Theories and Applications of Data Science; Algorithms and Programming/Data Structures; Recommender Systems and Web Mining; Convex Optimization for Data Science; Data/Web Visualization; Probability and Statistics. Numerical Analysis; Systems for Scalable Analytics; Modeling and ML; Advanced Algorithms; Deep Learning

EXPERIENCE

Synthetic Biology Knowledge System, San Diego Supercomputer Center February 2021 – Present
Research Intern
- Maintainer for **Topic Modeling** trained on 1000+ academic articles: achieved high topic interpretation score rated by domain experts. Results were used to find similar articles and build **author-article network**.
- Organized Text Parsing code (**XML parsing**) and **built customized Docker Containers** for different tasks.
- Built an interactive author-articles graph network (technical cluster & ethics cluster) using **Neo4j (graph database)**

Halıcıoğlu Data Science Institute, UC San Diego March 2020 – present
Instructional Assistant
- IA for DSC 40A (Math and Theories for Data Science & Machine Learning); DSC10 (Principles of Data Science)
- Help with structuring the class and holding **office hours**, **debugging** in programming assignments on Gradescope

CNS Laboratory of Memory and Brain January 2020 – June 2021
Undergraduate Research Assistant
- **Data cleaning/image processing**: worked on fill in missing values using statistical method, automating image edits
- Brain **data visualization** and **statistical analysis**: plotting using **matplotlib**, experimented **correlations** with **PCA**
- Individual Research Project on White Matter Hyperintensity with **prediction to Alzheimer's Disease**
- Wrote a **data pipeline** for quantifying WMH and analysis for 50+ subjects using **Shell Script and Python**

China Telecom, Suzhou, China June 2019 – July 2019
Intern, Assistant to the Network Optimizer
- Worked with wireless network optimization including concepts like **coverage, band frequencies**, etc.
- Analyzed and cleaned user data using **Pandas and Python** to assist creating network solutions

PUBLICATIONS

McInnes, Bridget, **"Discovering Content through Text Mining for a Synthetic Biology Knowledge System"** ACS Synthetic Biology 2021 (submitted and waiting to be published)
- Contribution to the **Topic Modeling** parts
- Previous Paper: Monte, Jeanet, "**Synthetic Biology Knowledge System**", ACS Synthetic Biology 2021, 10, 9, 2276-2285

PROJECTS

White matter hyperintensities and their relationship to demographics, cognition, and news event memory
- Created an automated date pipeline with Python and Shell to extract WMH volumes from brain scans
- Found significant relationship when regressing news event memory age against WMH volumes

Faculty Information Retrieval System for the Halıcıoğlu Data Science Institute
- Topic Modeling and NLP to group faculty members based on their publications to serve industry needs

Web based Data Visualization & Advertisement
- A dashboard-like web-based data visualization developed from scratch using Html, JavaScript, and CSS

Interactive Synthetic Biology Author-Articles-Institutions Graph Network Building hosted on Neo4j Aura Cloud
- Graph relational database constructed to show a graph network for publications in Synthetic Biology

Satisfaction Classification on Hotel Reviews using TFIDF, Word2Vec, Logistic Regression
- A classification model that applies logistic regression with other NLP techniques to predict user satisfactions

Lipstick Popularity Market Analysis published as online data science blog
- Recommendation of lipsticks based on sales data from Taobao; received good feedbacks from readers

China Earthquake Frequency Analysis published as online data science blog
- Interactive heatmap generated by 10000+ data scraped from earthquake report website