Tongmeng (Tommy) Xie | Data Scientist

Tel: +44-7536359076 | Email: TommyXie_@outlook.com
MicroSite: Streamlit (tongmengxie-researches.streamlit.app)
Add: London, United Kingdom

EDUCATION & ACADEMIC ACHIEVEMENT

09/2022-09/2023
University College London, The Bartlett Centre for Advanced Spatial Analysis
Program: **(Msc) Urban Spatial Science (Spatial Data Science**)
Degree: **Merit (2:1)**

- Dissertation (08/2023): *Address Matching and Entity Extraction Across UK Properties Data*, which attracted attention in the field of Record Linkage and landed me on a research assistant contract with Department of Geography and Environment, LSE.

09/2018-06/2022
Chengdu University of Information Technology, College of Software Engineering
Program: **(BEng) Spatial Information and Digital Technology (**A sub-discipline of **Computer Science)**
Overall Average Score: **86.19%**

- Dissertation (06/2022): *A Poet Powered by Transformers*, fine-tuned a GPT-2 model with poetry dataset in ancient Chinese (A year before ChatGPT came into view).
- National Scholarship (12/2021)
- Top Scholarship at Chengdu University of Information Technology (1st of 77) (12/2019)
- 1st-class Prize, English Copywriting of "Tongyi Cup" International Exhibition (1st of 3000) (12/2019)
- Top Prize, National English Competition for College Students (Top 1%) (06/2019)

INDUSTRIAL EXPERIENCE

09/2023-12/2023
***Data Analyst,* Albany Beck Consulting**

- Full-stack development of an event-management system with Object-Oriented Programming in Python
- Did risk analysis in synthesised FDI regarding UK-EU interconnectedness
- Research into carbon emission in auto sector using analysis tools like Scikit-learn and StatsModels, calculated Value at Risk and predicted future trends using Monte-Carlo Simulation

06/2021-11/2021
***Algorithm Engineer Intern*, Aplux Intelligent Technology**

- Developed demo app for industrial detection on AidLux platform
- Trained U-Net neural network with ResNet backbone to detect defects in industrial wood, Developed an industrial app to detect fabric defects based on Tianchi Big Data Competition project
- Developed apps using NLP for AidLux: Developed an app to generate ancient Chinese texts, Developed an app to punctuate sentences and classify texts using NLTK and Jieba

SKILLS

Data Analysis and Engineering Skills

- Extra Large Data Analysis: Handling out-of-core computation with batching, scheduling, streaming and lazy computation.
- Pipeline Building and Optimisation: Using profiling, monitoring and distributed computing to reduce latency in ETL pipelines.
- Data Analysis and Mining: Experience with Pandas, R, Excel, PowerBI; Data mining using Beautifulsoup4 and Regular Expression.
- Database and Warehousing: Proficient in SQL, Postgres; Hands-on experience with AWS; Structured database development and data warehousing.
- Big Data Engineering: Utilising distributed frameworks such as Dask and Ray for big data solutions.
- Visualisation: Skills in Matplotlib, Seaborn, interactive mapping, and deployment on micro-sites.

Natural Language Processing (NLP) and Machine Learning

- NLP Techniques: Record linkage, Named Entity Recognition with spaCy, Programmatic labelling with weak learning, Data augmentation via weak-supervision ML techniques.
- Machine Learning Pipelines: Building ML pipelines with NumPy, Pandas, scikit-learn, TensorFlow; Large Language Models (LLM) fine-tuning including prompt-tuning, parameter-efficient tuning; LLM inference such as zero-shot and few-shot learning.

Spatial and Quantitative Analysis

- Spatial Analysis: Proficiency with R, Python, QGIS, GEE for spatial analysis.
- Quantitative Modelling: Knowledge in Linear Algebra, Calculus, Statistics; Predictive modelling including time series analysis, Spatial Interaction Model, Network analysis, LULC models.
- Causal Inference: Regression Discontinuity Design (RDD) and Difference-in-differences (DID).

Tongmeng (Tommy) Xie | Data Scientist

Programming and Software Development Skills
- Python Programming: Includes Object-Oriented Programming (OOP), data structures/collections, functions, modules/libraries, Exception & File Handling, Functional programming, debugging, logging, Unit Testing (TDD), Multithreading, Closures, Decorators, Iterators,
- Tools and Frameworks: NumPy, Pandas, Matplotlib, Power BI, scikit-learn, TensorFlow.

Project Management and Agile Methodologies
- Software Development Life Cycle (SDLC), Agile methodologies, DevOps practices.
- Project Management Tools: Proficient in Jira and Confluence.

ACADEMIC ACTIVITIES

09/2023-01/2024
***Research Assistant,* Marshallian Theories In A Historical Context: Matching Across Censuses and Patent Data**
- Performed the memory-hungry task of linking between historical census records (1851-1911, UK & Wales) on HPC, achieved 45% match rate, a significant improvement on SOTA
- Innovatively replacing expensive hand-labelling with iterative learning techniques like semi-supervision to acquire initial labels automatically, then reduced label noises with ensemble learning, largely saved cost for crowd-sourcing and manual labour
- Extracted patent author using Named Entity Recognition (NER) building on basic rules
- Designed a learning framework containing Machine Learning (ML) base learners, e.g., Random Forest, XGBoost, CatBoost and a higher synthesising neural network, efficiently handled label noise

05/2023-08/2023
Dissertation: Address Matching and Entity Extraction Across UK Properties Data
- Operated matching between PPD and EPC (each of them with 2*10^7 records) using differently structured addresses
- Used PySpark framework and Parquet file format to handle large-scale data persistence and paralleling
- Ensembled Deterministic, Probabilistic, Distance-based and Natural Language Processing (NLP) approaches for address matching
- Achieved a 5% raise in matching rate compared to the SOTA hand-made matching rules.

03/2023-05/2023
Agent-based Modelling (Coursework): Evacuation Zoning Strategy in a Multi-modal Transportation Case
- Drafted an Overview, Design concepts & Details (ODD) Description of a disaster evacuation scenario based on San Macro,Venice
- Modelled traffic considering multi-modal dynamics of automobiles, pedestrians and motorcycles, using NetLogo

02/2023-04/2023
Agent-based Modelling (Coursework): Resilience of London Underground from an Urban Simulation Perspective
- Calculated Centrality measures of underground stations in London. Simulated what-if scenarios to measure station importance
- Modelled hypothetical conditions (new airport in area, increase of travel cost) using Spatial Interaction Models and visualisation

11/2022-01/2023
Quantitative Method (Course Work): Exploration and Quantification of Factors Influencing KS4 Performance in England
- Quantitatively modelled urban outcomes using statistical models e.g., OLS and ML models e.g., K-means clustering
- Completed data pipeline using Jupyter: Data scraping, pre-processing, exploring, modelling and evaluating
- Concluded that a 1% increase in the ratio of not disadvantaged children in a local authority leads to a 0.25 increase in average attainment 8 score, among other measured outcomes

07/2021-08/2021
***Team Leader,* National College Student Curling Artificial Intelligence (AI) Challenge—Digital Curling Competition**
- Designed an AI programme based on reinforcement learning to compete with nation-wide teams in a virtual curling stadium, finished in the top 8
- Experimented with deep Q-learning and Actor-critic algorithms using python
- Found one deterministic policy written in C++ to be performing stably well in the virtual stadium

11/2020-01/2021
***Data Engineer, Big Data Systems and Technologies (*Coursework):** Utilised Apache Hadoop distributed computing framework to perform sentiment analysis using Alibaba comment data
11/2020-12/2020
Geographic Information System (GIS) Development Project: Developed a web-based street map system, added display feature

EXAMINATIONS & ADDITIONAL INFORMATION

GRE: 330 (Verbal: 164; Quant: 166), Writing: 3.5 **Languages**: English (proficient), Japanese (intermediate), Chinese (native)
Patents: A Computility Device. This device re-uses computational devices like phone chips
Extra-curricular: 1. ***Tutorship with Students with Disadvantaged and Under-represented Backgrounds***, Action Tutoring. 2. ***Membership and activities in Student Action Against Homelessness Society***, UCL. Badminton, Swimming, Calligraphy, Chess.