JINGJING LIN (She/Her)

+1 202.460.4927 | isjingjing.lin@gmail.com | Arlington, VA 22202 | ⚙ github.com/JJJJJingL 🔗 LinkedIn-imjingjinglin

EDUCATION

Georgetown University — 08/2018 – 05/2020, Washington D.C.
- Master of Science: Data Science and Analytics, GPA 3.5/4.0
- Peer lead mentor, advised mentors and ~70 graduate students

University Of Manchester — 09/2015 – 12/2016, Manchester, UK
- Master of Science: Management and Information Systems, GPA 3.3/4.0

Tianjin Polytechnic University — 09/2011 – 06/2015, Tianjin, China
- Bachelor of Engineering: Software Engineering, GPA 85/100 | Minor: Finance, GPA 88/100
- 2014 Presidential First-Class (top 3%), 2013 Second-Class (top 5%) Scholarship; Outstanding Graduate (top 2%)

SKILLSET

Programming: Python (Scikit-learn, pandas), **R** (dplyr, glmnet), **SQL**, VBA – Demo, Java (JDBC), HTML, CSS, Shell

Machine Learning: Regression (Linear/Logistic, Lasso), Clustering (K-Means, Hierarchical), Classification (Boosting, Random Forest, SVM), Deep Learning (CNN, RNN), Natural Language Processing (NLP), Dimension Reduction (PCA)

Statistics: Probability, Distribution, Sampling, Hypothesis Testing, Bayes Theorem, Correlation, A/B testing

Cloud Computing: AWS (EMR, S3, Hadoop, MapReduce, Spark, Git); Google Cloud (BigQuery, storage buckets)

Visualization & Tools: Tableau, Plotly, Matplotlib and ggplot2; MySQL, Jupyter, Google Analytics (Certificated)

EXPERIENCE

Georgetown University Graduate School | *Higher Education* — 08/2020 – Present, Washington D.C.
Data Science Researcher
- Collected 100+ vaccines development progresses from scientific reports, merged data using pandas and NLTK
- Built a map-based Tableau dashboard to quantify and visualize textual processes with geological information
- Generating weekly reports and embedded results in the website using HTML and CSS

Center for Security & Emerging Technology, Georgetown University | *Higher Education* — 09/2019 – 12/2019, Washington D.C.
Data Science Research Assistant
- Accessed data from 3 AI publication databases (130+ million rows of 14+ GB) using BigQuery (SQL) in GCP
- Merged database A, B by DOIs and metadata (e.g. keywords), built a dictionary from metadata by bag-of-words
- Found the top 50 matches in B for each article in A by matching TF-IDF distance and the year of publication, created the distribution of the longest substring sizes in each match of the top 50 articles
- Increased matching rates to ~20% in articles without DOIs by running text similarity algorithms using the highest similarity score, based on the threshold (95%) from the scores distribution in matched articles with DOIs

Unilever - Dollar Shave Club | *Consumer Goods* — 06/2019 – 08/2019, Los Angeles, CA
Marketing Technology Intern
- Sole analyst responsible for analyzing and optimizing the current manual-operated 20+ spreadsheets with 200+ tags of Urchin Tracking Module (UTM) information management
- Reduced parameters setting time by 90% for marketing team, and implementing time by 33% for data systems team by establishing a new automated tags generator and re-designing the dashboard using VBA, SQL and Macros
- Delivered a new UTM solution to teams (~20 members) independently, designed a long-term operation plan
- Investigated and organized existing marketing campaigns and supply chain, to create a business proposal for 'DSC X Military' to build connections with military communities

WST Consulting (Startup) | *Education Consulting* — 09/2017 – 04/2018, Shanghai, China
Research Analyst
- Gathered information on target firms' finance and development strategy to generate biweekly reports, created 4 chapters of guidebook on recruitment programs of finance, consulting, data and tech fields across global markets
- Led resource management efforts to create and restructure marketing materials that yielded 50% increase (from ~3000 to ~4500) in average view count of over 15 supported articles on WeChat platform
- Supervised an intern and 2 junior colleagues on document research methods and writing materials revision

PROJECTS (More projects accessible on GitHub)

Data Science: Where Should You Live for Your Health (Demo, Code)
- Acquired data (~20k rows) through API after scoping the project and translating it into an analytical problem
- Assessed data by performing data wrangling and normalization with pandas, posed 3 hypotheses
- Architected multiple ML models (clustering, association rule mining and regression) on factors (air and water quality) to validate hypotheses, investigated the cancer trend based on results in Tableau and network analysis

AWS - Data Engineering: Top Comment Identification in Reddit (Code)
- Accessed and loaded datasets of Reddit comments (~500GB) in JSON from S3 and preprocessed data, i.e. cleaning (handling missing and inconsistent values), discretization, transformation and sampling, using PySpark in EMR
- Performed EDA with Spark SQL; created features in numeric (text-length) and categorized (scores) variables
- Conducted features encoding through MLlib; modeled a top comment identifier with logistic regression through machine learning pipeline, achieving the accuracy of 90%