

JINGJING LIN

isjingjing.lin@gmail.com | Arlington, VA 22202 | 202.460.4927

 [JJJJJingL](#)  [jingjingl.georgetown.domains](#)  [imjingjinglin](#)

SKILLSET

Programming	Python(sklearn, pandas), R(dplyr, glmnet), SQL, VBA(Excel-Macro), JAVA, HTML, CSS, C
Machine Learning	Regression, Bayesian, Ensemble, Decision Tree, Clustering, Deep Learning (CNN, RNN), NLP
Visualization	Tableau, Plotly, Matplotlib, ggplot2 and R-markdown
Cloud Computing	AWS (EMR, S3, Hadoop, MapReduce, Spark, git); Google Cloud (BigQuery, storage buckets)
Database & Tools	RDBMS: MySQL (JDBC) and Access; Command Line, Jupyter notebooks

EDUCATION

Georgetown University, USA	– Master of Science, Data Science and Analytics	2018 – 2020
University of Manchester, UK	– Master of Science, Management and Information Systems	2015 – 2016
Tianjin Polytechnic University, China	– Bachelor of Engineering, Software Engineering	2011 – 2015
Tianjin Polytechnic University, China	– Bachelor of Economics, Finance	2011 – 2015

EXPERIENCE

Data Science Development Engineer – Georgetown University, Washington, D.C.	Aug 2020 – Present
<ul style="list-style-type: none">Developing methods to track news and scientific papers related to COVID-19 using APIs (web-scraping)Building data-oriented features (e.g. visualizations) to explain the scientific progress in the fight against COVID-19	




Data Science Research Assistant – The Center for Security and Emerging Technology of Georgetown University, Washington, D.C.	Sep – Dec 2019
<ul style="list-style-type: none">Performed exploratory data analysis (EDA) on academic publication datasets to characterize tech fields in Artificial Intelligence through BigQuery, storage buckets, and virtual machines in Google Cloud ConsoleConducted textual analysis, including converting bags-of-words, vectorizing tf-idf and running text similarity algorithms, to increase matching rates across academic publication databases	

Marketing Technology Intern – Dollar Shave Club Inc., Los Angeles, CA	Jun – Aug 2019
<ul style="list-style-type: none">Developed an Urchin Tracking Module (UTM) parameters generator tool independently to manage Ads campaign information using VBA and SQL; designed a plan for long term maintenance and operations across the companyImplemented marketing integrations in tag management systems from Google Analytics to Adobe AnalyticsCreated a business proposal for ‘DSC x Military’ to build connections with military communities	

Research Analyst – Wall Street Tequila Consulting Inc., Shanghai, China	Sep 2017 – Apr 2018
<ul style="list-style-type: none">Investigated the trend on target firms’ recruitment plans and strategies to generate guides and periodical reportsCreated writing materials by restructuring resources to support marketing team (yielded 50% growth in average view count of 15 articles on WeChat platform) and consulting team (developing speech drafts and slides)	

Software Dev Engineer Intern – ChinaSoft International Ltd., Tianjin, China	Summers, 2012 – 2015
<ul style="list-style-type: none">Designed and built UI, database and prototype for 4 systems: [1] ‘Dieting Assistant’ Fitness System (2015), [2] Veterinary center management system (2014), [3] Online shopping website (2013), [4] Static social website (2012) with Java, HTML, CSS and MySQL (JDBC) for 3 consecutive summersDocumented feasibility analysis reports and project development plans; delivered final presentations	

PROJECTS

-  **Massive Data: Top Comment Identification in Reddit** Apr – May 2019
 - Accessed large datasets of Reddit comments(~500GB) in JSON and preprocessed data using PySpark in EMR
 - Performed EDA with Spark SQL; created features in numeric (text-length) and categorized (e.g. score) variables
 - Built “pinned” comment identifier by applying new features to logistic regression through machine learning pipeline
-  **NLP: IMDB Rating Prediction by Modeling Movie Scripts** Mar – Apr 2019
 - Collected ~1300 film scripts from 22 genres and their IMDB ratings, preprocessed datasets with NLTK
 - Calculated and vectorized features, such as tf-idf, the mean number of words per sentence and “pos tag” frequency
 - Trained linear regression and Random Forest models with different feature combinations with sklearn; compared the two models using Pearson’s r and demonstrated the performance of Random Forest reaching an accuracy of ~85%
-  **Data Analytics: Where Should You Live for Your Health** Sep – Dec 2018
 - Acquired datasets through API and performed data wrangling (~20k rows) to classify water quality data with pandas
 - Implemented clustering (e.g. k-means) and association rule mining analysis, visualized them by Tableau and Plotly
 - Applied hypothesis testing on cancer rates by using linear regression and classifiers e.g. KNN, Naïve Bayes, SVM