LUYAO PENG, Ph.D.

Email: luyao.peng@email.ucr.edu Website: pengluyaoyao.github.io Phone: (619)-313-3133

Data scientist in NLP (natural language processing) with experience in text generation, general language understanding evaluation and chatbot, and background in statistics (Ph.D.) and linguistics (M.A.).

WORK EXPERIENCE

Data Scientist II, Artificial Intelligence & Machine Learning Team, ACT, Inc.
August 2019 — July 2020

- Developed and led research projects on scalable deep-and-wide learning with NLP models
- Improved and deployed machine/deep learning algorithms in multiple NLP applications and products
- Model training, validation, and product development of the CRASE+ automated scoring engine

Statistical Consultant, University of California, Riverside
September 2015 — November 2016

- Led statistics workshops with topics on data mining, programming, and reproducible research
- Provided consultations on empirical research methods in various applied disciplines

Research Intern (Machine Learning), CTB McGraw-Hill Education
June 2015 — August 2015

- Conducted machine learning research on automated essay scoring and forensic analysis
- Developed and deployed online visualization application for fraudulent detection

Teaching Assistant, University of California, Riverside
September 2016 —June 2019

- Led teaching sessions of statistics courses (probability theory, sampling, and statistical inference).

TECHNICAL SKILLS

Programing	Python, R, Shell, SQL, Git
Deep Learning for NLP	TensorFlow, PyTorch, Transformers, fairseq, ParlAI
NLP Model	Deep and Wide, RNN-based model, Encoder-decoder model, BERT, GPT
Machine Learning Model	Regression, Classification, Clustering, feature engineering, topic modeling
Big Data Processing	PySpark, HPCC, AWS
Statistics	Mixed-effects Model, Multivariate Model, Gaussian Process, Probability
Visualization	Bokeh, Seaborn, Matplotlib, Shiny, Lattice, ggplot2

SELECTED PUBLICATIONS AND WORKING PAPERS

- Luyao Peng, Saad Khan. Gaussian Process Deep-and-Wide Regression Model. *The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP2020) (under review).*

- Luyao Peng, Saad Khan. Scalable Gaussian Process Deep-and-Wide Model for Classification. *The 28th International Conference on Computational Linguistics* (*COLING2020) (under review).*

- Luyao Peng, Sandip Sinharay. Using Linear Mixed-effects Models to Detect Fraudulent Erasures at an Aggregate Level. *Journal of Educational Behavioral Statistics (under review).*

- Luyao Peng, Subir Ghosh. An Algorithmic Construction of All Unbiased Estimators of Variance Components in Linear Mixed Effects Models. *Joint Statistical Meeting by American Statistical Association. 2019.*

- Subir Ghosh, Li Guo, Luyao Peng. 2018. Variance component estimators OPE, NOPE and AOPE in linear mixed effects models. *Australian & New Zealand Journal of Statistics 60*(4), 481-505.

RESEARCH PROJECTS

Gaussian Process Deep-and-Wide Model

- Developed gaussian process deep and wide model for prediction and classification
- Applied the model to the ASAP essay data and large-scale MIMIC3 medical data

Bert Summarization and Generation of Chinese Language

- Implemented existing research on Bert Generation on Chinese language
- Extractively summarized Chinese stories using mixture gaussian and k-means clustering

Neural Text Generation

- Trained encoder-decoder model (transformer network models) to generate storylines given keywords
- Trained a different encoder-decoder model to generate story given storylines

Fraudulent Response Detection in Automated Essay Scoring

- Applied Kernel Principal Component Analysis (KPCA) and Support Vector Machine (SVM) to detect abnormal essays in automated essay scoring examination and cheating erasures in forensic analysis
- https://kpca-outlier-detection.shinyapps.io/Shiny/

PACKAGES

"MMeM" (**Multivariate Mixed-effects Model**)

- Developed and maintained an R package for modeling multivariate mixed-effects using REML and Henderson3 methods.
- https://CRAN.R-project.org/package=MMeM

"regrrr" (**Toolkit for Compiling and Visualizing Regression Results**)

- Coauthored an R package for regression result reporting, hypothesis testing, and visualization.
- https://CRAN.R-project.org/package=regrrr

EDUCATION

University of California, Riverside 2019
Ph.D. in Applied Statistics.

Fellow, The Data Incubator (TDI) 2018
Trained with Data Science skills in machine learning toolkit, web scraping, SQL, mapreduce, Natural Language Processing, Spark and Tensorflow

University of California, Riverside 2014
M.A. in Educational Psychology: Quantitative Research Methods

Beijing Language and Culture University 2012
M.A. in Linguistics

Capital Normal University 2009
B.A. in English Language and Literature

CERTIFICATES AND ACTIVITIES

- Databricks Certification for Apache Spark, 2020.
- Co-founding Vice President, Data Science Club, University of California, Riverside, 2016-2019.
- R and Spark: Tools for Data Science Workflows, NISS, 2017.
- Graduate Division Fellowship Award (Excellent Student), University of California, Riverside, 2013.