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
- 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 abnormal behavior detection in automated essay scoring and forensic analysis
- Developed and deployed an 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)

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, Jianqin Wang. Scalable Gaussian Process Deep-and-Wide Model. *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 (revise and resubmit).*

- 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.

- Luyao Peng, Raghuveer Kanneganti. Statistical High-dimensional Outlier Detection Methods to Identify Abnormal Responses in Automated Scoring. *National Council of Measurement in Education Annual Meeting, 2016.*

- Luyao Peng. Deterministic, Gated IRT Model for Continuous Probability of Item Cheating. *National Council of Measurement in Education Annual Meeting, 2015.*

- Gregory Palardy, Luyao Peng. 2015. The Effects of Including Summer on Value-added Assessments of Teachers and Schools. Education Policy Analysis Archives, 23(92), 1-26.

RESEARCH PROJECTS

Gaussian Process Deep-and-Wide Model [link]
- Developed scalable 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 latest research on BERT Generation of Chinese language
- Extractive summarization of Chinese stories using mixture Gaussian and k-means clustering

Chatbot Using GPT and GPT2 [link]
- Implemented GPT and GPT2 double headed language model to build a simple chatbot with personalization.

Fraudulent Response Detection in Automated Essay Scoring [link]
- 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

PACKAGES

"GaussianProcessClassificationModel" [link]
- Developed gaussian process classification model and example using tensorflow-probability.

"MMeM" (**Multivariate Mixed-effects Model**) [link]
- Developed and maintained an R package for modeling multivariate mixed-effects using REML and Henderson3 methods.

"regrrr" (**Toolkit for Compiling and Visualizing Regression Results**) [link]
- Co-developed an R package for regression result reporting, hypothesis testing, and visualization.

EDUCATION

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

The Data Incubator (TDI) 2018
Fellow in Data Science. Trained with 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: second language testing and acquisition

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

ACTIVITIES, AWARDS, AND CERTIFICATES

- Co-founding Vice President, Data Science Club, University of California, Riverside, 2016-2019.
- Innovation and Entrepreneurship Award, UCR Office of Technology Partnerships, 2019.
- Databricks Certification for Apache Spark, 2020.

- R and Spark Certification: Tools for Data Science Workflows, NISS, 2017.
- Excellent Ph.D. Student Fellowship, Graduate Division, University of California, Riverside, 2013.
- Excellent Paper Award, Second Language Acquisition Forum, Peking University. 2012.

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