Zhengtong Pan

Burlingame, CA, 94010 | (925)-434-6358 | pztcookie@gmail.com | Linkedin: www.linkedin.com/in/zhengtong-pan

SUMMARY

Solid Data Science skills with 2 years of project experience in question formulation, data collection and cleaning, data visualization, statistical inference, predictive modeling, and decision making.

EDUCATION

University of California, Santa Cruz **Santa Clara, CA**
Master of Natural Language Processing *Sept 2022 ~Aug 2023*
University of California, Davis **Davis, CA**
Bachelor of Science in Data Science, Overall GPA: 3.57 *Graduation in June 2021*

SKILLS

Programming Languages and Tools

Python (Numpy, Pandas, Scikit-Learn), SQL, Scala, Spark, Keras, Natural Language Toolkit, Gensim, AWS

Machine Learning

Supervised Learning	Classical (Linear, Logistics) & Penalized (Lasso, Ridge) Regression (LR), Support-Vector Machines (SVM), Multinomial Naive Bayes (MNB), Decision Tree, Random Forest, Gradient Boosting Decision Tree, K Nearest Neighbors, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN)
Self-supervised learning	Zero-Shot Learning (ZSL: BERT, BART)
Unsupervised Learning	K-means, Latent Dirichlet Allocation (LDA), Alternating Least Squares
Natural Language Processing	Text Preprocessing (Stopwords Removal, Tokenization, Stemming, TF-IDF), Word Embedding (BOW, Word2Vec, GloVe), Sentiment Analysis, Topic Modeling, Long Text Classification, Named Entity Recognition (NER), Transformers
Dimensionality Reduction	Principal Component Analysis, Non-Negative Matrix Factorization

Statistical Analysis

Hypothesis Testing, Experimental Design, A/B Testing, Time Series Analysis

WORK EXPERIENCES

Wissee **Seattle Washington (Remote)**
NLP/Machine Learning Research Assistant *Sept. 2021 - June 2022*

· Automated multilabel business event classification for fashion businesses based on influencers' social media posts to guide traditional businesses to make data-driven marketing strategies.
· Optuna auto-tuning to develop **CNN** with f1-score above 0.80 in classification for online and offline business events.
· Automating sentiment classification for fashion influencers' social media posts with a 0.75 f1-score.
· Developing rule-based purchase intention classification for sales and non-sales fashion influencers' social media posts.

Haystack Search **Los Angeles, CA (Remote)**
Data Scientist (Part-time) *Jul. 2021 - May 2022*

· Automated the classification for local retailer products to build trust between offline local businesses and consumers.
· Developed **Multinomial Naive Bayes** based on products' titles and abstracts, with an average of 0.85 accuracies.

Department of Computer Science, University of California, Davis **Davis, CA (Remote)**
NLP Researcher Sept. 2020 - Sept. 2021
Publication(ICDMAI): Ontology-driven Scientific Literature Classification using Clustering and Self-Supervised Learning

· Built hierarchical document classifiers to classify 52,000 unlabelled research papers into emerging tech fields to eliminate the cost, risk, and inconsistency of manually processing rapidly growing volumes of documents.
· Collected labels from conferences' "Call For Paper" pages and constructed an **Ontology** to build hierarchical labels.
· Automatically assigned labels using **ZSL** with a 0.80 confidence score to eliminate the cost and error in manual labeling.
· Preprocessed the abstracts of Google Scholar papers by extracting n-grams and removing a 90% non-useful corpora.
· Vectorized text with **BOW, Word2vec, GloVe**, and conducted labeling verification with **Agglomerative Clustering**.
· Resampled imbalanced data, and developed **RNN** above 0.90 f1-score at each granularity level.

PROJECT

Sentiment Analysis and Topic Modeling on E-commerce Customer Reviews (Python)

· Performed sentiment analysis to reveal polarity in customer reviews, extracted latent topics based on the sentiment results to gauge hidden customer demands and concerns, and offered brand monitoring to e-commerce retailers.
· Preprocessed customer reviews utilizing stopwords removal, tokenization, stemming, and vectorization with **TF-IDF**.
· Performed topic modeling and trained **K-means Clustering**, **Latent Dirichlet Analysis**, and **Non-negative Matrix Factorization** to discover hidden semantic structures in customer reviews.

Movie Recommendation Engine Development (Apache Spark)

· Developed a movie recommendation engine in Spark to engage movie lovers by delivering personalized movie content.
· Built data ETL pipeline to analyze movie ratings with **Spark SQL** and monitored the performance via **Spark UI** on **AWS**.
· Performed collaborative filtering with **Alternating Least Square** to recommend movies based on latent movie factors.