Pham, Minh Quang

Data Scientist

github.com/tisu19021997
quangphamm@gmail.com
+84 769009285

A data-driven scientist with a product-focused mindset. Experienced in designing and implementing large-scale customer data platform. Always strive to leverage data science and machine learning to tackle complex problems and deliver easily accessible solutions to customers. Particular interest in NLP and LLMs and their application in business settings.

Work Experience

Data Scientist

Nov 2020 - Aug 2023
Center of Applied Data Science, FPT

Developed large-scale in-house marketing customer data platform

- Leveraged machine learning and natural language processing algorithms to extract, and enrich customer information.
- Engineered a production-grade, high-speed identity resolution algorithm in PySpark for over XX million customers across 5 diverse data sources.
- Led the product initiation and development team to design data models, product features, interfaces, and visualizations.

Long Chau customers potential health issues prediction

- Utilized NLP models (BERTs family) to predict health issues using customer's pharmaceutical transactions. Achieved 85% Jaccard similarity compared to human annotation. Further enhanced by combining web scraping and ChatGPT API.
- Business-oriented customer segmentation with RFM analysis and K-Means algorithm.

Others

- Implemented and maintained multiple daily data pipelines for XX thousands of transactions.
- Developed insightful PowerBI dashboards for monitoring sales and customer return rate.
- Trained an LSTM model with 90% accuracy in predicting customers' gender from the full names.

Techinical Experience

Kaggle NLP Competition - Feedback Prize 3 kaggle.com/quangphm

- Fine-tuned SOTA transformer models like RoBERTA, DeBERTa, GPT-2. Ensembled 12 language models with RAPIDS SVR. Personal best score 0.4373 compared to 1st rank 0.4333.

LLMs fine-tune and application github repository

- Built high-quality 15K Vietnamese instruction-following synthetic dataset using ChatGPT with the help of Selenium agent to reduce to cost of calling OpenAI API.
- Leveraged HuggingFace (HF), PEFT, accelerate libraries and Kaggle free T4s to fine-tune 7B and 13B Llama-2 models.
- Built web-based RAG application on resumes to allow recruiters to chat with candidates' resumes (w.i.p). Tech stack: HF, LlamaIndex, NextJS, FastAPI, PostgreSQL.

Education

Bachelor of Engineer, Computer Science

International University, Vietnam National University | 2019

Awards and Achievements

Most Valuable Player Q1/2021

Employee of the year 2021

Publications

Incremental SVD-based Collaborative Filtering Enhanced with Diversity for Personalized Recommendation

Accepted at ICCI 2020

Tech stack

pytorch, huggingface, llamaindex
pyspark, scikit-learn, pandas, numpy
postgresql, javascript, figma