



zhoumi

EDUCATION

CORNELL UNIVERSITY

M.Eng. in Computer Science, GPA: 4.035/4.3

Ithaca, NY

Jan 2019 - Aug 2020

GEORGIA INSTITUTE OF TECHNOLOGY

B.S. in Computer Science & B.S. in Mathematics, Highest Honors, GPA: 3.87 /4.0

Atlanta, GA

Jan 2014 - Dec 2017

EXPERIENCE

POP SOCIAL (WEB3 STARTUP)

San Jose, CA

Machine Learning Engineer, AI Team

Mar 2023 - Oct 2023

- Participated in the design and development of ELT pipelines that can support both batch and stream processing, using **Kafka, Spark Streaming, Hbase, Pinecone**, and **Airflow** as the orchestrator to digest user engagement data (click/add2cart/purchase) generated in recommendation log systems
- Enriched the rank stage of the recommendation system by inserting a transformer layer between the embedding layer and MLP to learn better representations for users' clicked items by capturing underlying sequential signals using **Keras**, attaining an AUC of 0.789 in offline experiments and gaining an improvement of 4.57% in CTR in online A/B tests
- Worked closely with product managers to implement a real-time auction system in **Gin** for live-stream shopping, providing 80k+ crypto merchants with highly consistent facilities to add/delete, update and release their digital products
- Generated and deployed daily & hourly batch data pipelines capable of ingesting 100+ query tables from different databases into **PostgreSQL** with **Beam & PyDolphinScheduler**
- Revamped monitoring dashboards & alerts with **Prometheus & Grafana** to better showcase ML model training logs and performance metrics for canary tests of model deployment

THE MICHAELS COMPANIES, INC.

Irving, TX

Software Development Engineer Back-End 3, Search & Recommendation Team

Nov 2020 - Nov 2022

Query Understanding

- Developed a query spell checker based on **GPT-2** in the **Noisy Channel** framework to tackle real-word errors, achieving an F1 score of 0.593 for detection and 0.508 for correction on full streaming A/B testing
- Designed and engineered an extreme multi-label classifier based on **CNN-RNN** with soft F1 loss using **Theano, SciPy, Gensim**, and **NLTK** for query enrichment, achieving an F1 score of 0.729, and PR@1, PR@3, PR@10 scores of 0.766, 0.817 and 0.878 respectively as classifying search queries into subsets of over 30k pre-defined leaf categories
- Developed query understanding APIs with **NLTK, Gensim**, and **TextBlob** to produce enhanced queries for the search & recommendation engine, including query embedding, POS parsing and tagging, synonyms, and spelling variant generation
- Inaugurated a **DSSM-like** matching algorithm for semantic product search using **Keras** and **Scikit-Learn**, where a pre-trained **DistilBERT** was employed as the encoder and element-wise multiplication & subtraction was applied to the embedding matrices of query and product to calculate their relevance, outperforming **BM25** by 5.62% in MMR

Product Profile Establishment

- Incorporated session splitter and multithreading into a basic web crawler with **Selenium, Concurrent**, and **BeautifulSoup**, leading to a 12% higher scraping accuracy and 40x faster tracking information retrieval
- Implemented rule-based parsers with **spaCy, Pattern**, and **YAKE** to formulate tracking product fields including title, description, metadata, brand, numeric, and product search terms (queries from click logs) for 500k+ instances
- Created a taxonomy-enhanced classification channel built upon the fine-tuned **Roberta-Large-MNLI** textual entailment model using **PyTorch** and **Transformers** for product tagging, attaining an MMR score of 0.692 on internal testing sets
- Established a hierarchical multi-label classifier based on a dual-encoder architecture: a pre-trained **BERT-base-uncased** model for product encoder and a position-enhanced **GNN** for label encoder using **PyTorch** and **AllenNLP** to generate product profiles, achieving example-F1, P@1, P@3 scores of 0.816, 0.894 and 0.806 respectively in offline experiments
- Cooperated with data scientists to develop NLU micro-services, including sentiment analysis, text annotation, brand entity recognition, and content classification, and to containerize the services with **AWS Elastic Load Balancer** and **Auto Scaling**, providing highly scalable text mining services to 11 production environments
- Collaborated with ML engineers to design and develop RESTful APIs on top of trained NLP models in **Flask** and **Django** frameworks to automate content moderation, including text, image, and audio toxic content detection

TECHNICAL SKILLS

- | | |
|-------------------------------------|---|
| • Programming Languages: | Python, Golang, Java, Julia |
| • ML & LLM: | PyTorch, Keras, Scikit-Learn, Gensim, LangChain |
| • Data Mining & NLP: | NumPy, Pandas, YAKE, TextBlob, SciPy, NLTK, spaCy, AllenNLP |
| • Framework & Container: | Flask & Django, Gin, Kafka, Spark Streaming, Kubernetes |
| • Database & Cache: | PostgreSQL, MongoDB, Redis, HBase, Pinecone |