Saurabh Khanal

Herndon, VA, US | +1 (571) 723 0095 | skhanal8@gmu.edu
https://github.com/SaurabhK24 | https://www.linkedin.com/in/saurabh-khanal-516516171 |

EDUCATION

George Mason University **Fairfax, VA**
B.S in Computer Science
Minor in Data Science & Statistics

WORK EXPERIENCE

Software Engineer - SparkSoft **June 2023 – Current**

- Spearheaded transition from manual database backup process to an automated Jenkins-integrated pipeline, reducing backup time for dynamo-db backup tickets by 50% and accelerating delivery for EPS team
- Developed secure Spring Boot APIs integrating AWS secrets Manager enabling management of application secrets with defined operations. Implemented optimized service layer interactions to improve system reliability
- Resolved backend service API call failures by analyzing Splunk logs, F5 iRules, Nginx config files, and running curl commands to verify endpoints leading to enhanced system scalability and performance
- Integrated OAuth authentication for seamless user sign up/log in into hosted client applications

Software Engineering Intern - HeadStarter Accelerator **June 2022 - Aug 2022**

- Implemented web apps to production using React, AWS, and Rest APIs to open-source projects
- Built serverless APIs using AWS Lambda with 99% uptime and real-time DynamoDB data entry tables
- Improved resume parsing web application API response time's by 20% by reducing redundant database queries and caching reused data

ML Undergraduate Researcher - Georgetown University **May 2021 - Aug 2021**

- Engineered peak vegetation prediction with ML algorithms (random forest, gradient boosting), resulting in an exceptional 90% correlation with actual greenness levels/historic data
- Improved model's precision yield by over 15% by helping data transition from random forest model to gradient boosted model
- Overlooked extensive netCDF data, effectively handling 1TB+ datasets, a critical contribution to reliable predictions

PROJECTS

AutoLabz (https://www.autolabsz.com/) **May 2024 - Current**

- Engineered a hybrid car search with sparse-dense vectors using NLP and multiple LLM API's
- Utilized Sentence Transformers, BM25 encoder, and rerank to enhance search accuracy and relevance
- Architected a scalable RESTful API with Flask & Node along with Pinecone Indexes and LLMs function calling

Bi-Gram Language Modeling **June 2024 – Aug 2024**

- Developed a bigram language model for text prediction, utilizing NLTK to preprocess and tokenize large datasets, enhancing word pair prediction accuracy.
- Implemented Laplace smoothing to handle unseen word pairs, improving the model's generalization and reducing overfitting in sparse data environments.

Character Level Language Model **Aug 2024 – Current**

- Implemented an autoregressive character-level language model using PyTorch, featuring advanced architectures including Transformers with multi-head self-attention, GRUs, and MLPs with up to 4 layers and 384 neurons.
- Developed a flexible Python codebase supporting various model configurations, including dynamic adjustment of batch sizes, learning rates, and context lengths up to 64 characters for improved text generation quality.
- Optimized model training through techniques such as AdamW optimizer, learning rate decay, and gradient clipping, achieving efficient convergence and stable training for models with up to 11 million parameters.

SKILLS

Programming Languages / Tools : Java, C++, Python, JavaScript, R, Angular, React, PostGreSQL, MongoDB, Express, Node.js, TensorFlow, NumPy, Jenkins, Splunk, GitLab/GitHub, SpringBoot, REST APIs, Micro-services, AWS, Atlassian Jira, CI/CD
Interests: Option/Equity Trading, Competitive Basketball, Weight Lifting, Music, Traveling