SAIKIRAN PUVVADA

🌐 Portfolio   ⬛ puvvadasaikiran   ✉ saikirann.puvvada@gmail.com   in puvvadasaikiran   ⬛ +91-7658970090

SUMMARY

Data Science professional with 2.5 years of experience delivering successful end to end NLP/Speech Projects in US Healthcare. Great at understanding user needs and designing simple solutions to address complex problems, and communicating with empathy.
Currently Looking for opportunities as a Senior Data Scientist / Consultant / AI Product Manager

SKILLS

Programming/ Scripting Skills: Python, JAVA, Bash
Statistical Analysis: Descriptive Statistics, Hypothesis Testing
Database: SQL/MYSQL, MongoDB
Rapid Deployments: Gradio, FastAPI, Steamlit, Flask
Data Scraping and Analysis: Pandas, NumPy, BeautifulSoup, Selenium, ScraPy
Data Visualizations: Tableau, Excel, Matplotlib, Power BI, Illustrator
Platforms: Microsoft Office Suite, GCP, AWS
Version Controlling Tools: Git, BitBucket
Product Management Tools: Monday.com, Jira
Data Engineering Tools: Docker, Kubernetes, Kafka, Spark, Pub-Sub, AMPQ 0-9-1, Celery
UI/UX Prototyping: Figma, Adobe XD

EXPERIENCE

Senior Data Scientist (Speech/NLP) at Edvak Health [Apr 2023 : Present]

- **Real Time Conversation Capture Stack for US Healthcare**

  - Designed and Implemented scalable Conversation Capture Stack for EHR to reduce documentation burden on providers by 50% (Python, SocketIO, sklearn, Pytorch)
  - Led the deployment of the system from scratch, using Websockets to ensure scalability and reliability (Kubernetes, CI/CD, Docker)
  - Guided a 4 member team of junior data scientists in the modelling process, using advanced speech processing techniques and Large language Models(LLM) (SpeechBrain, PyAnnote, Clustering, Role identification, GPT4, OpenAI)
  - Successfully Integrated NLP/text dependent role identification techniques to improve on relative speaker confusion rate by 28% in a conversation

- **Front Desk call Automation for US Healthcare**

  - Designed and Implemented Call Automation for US Healthcare with VITS based Text to Speech Synthesis and LLM for appropriate response generation.
  - Designed scalable deployment strategy to handle patient queries over call in place of front desk staff.
  - Utilized LLM's Natural Language Undestanding and Natural Language generation capabilities for response generation using prompting techniques.
  - Utilized Large Action Models to perform actions based on patient's conversation flow appropriately.

- **Auto-Punctuation and Capitalization Model for US Healthcare**

  - Developed BERT based Auto-Punctuation and Capitalization Model for improving clinical document readability
  - Utilized Data Augumentation/Generation/Scraping techniques to tackle data scarcity (Pandas, Regex, Selenium)
  - Improved punctuations and capitalization performance by 24% over base model using re-defined modelling objective for in-domain transfer learning.

- **Managerial Responsibilities**

  - Co-ordinating with Data Science Team and Integration Teams to optimize the work and plan to align towards long term Organisational Business Objectives.
  - Managing and Guiding Junior Data Scientists in end to end project life cycle.
  - Communicating the with stakeholders to discuss plan/progress.
  - Managing and Planning Annotation Activities for Speech/NLP Usecases

Data Scientist (Speech/NLP) at Edvak Health [Dec 2021 : Apr 2023]

- **Automatic Speech Recognition(ASR) for US Healthcare**

  - Worked on developing Closed Language Model Automatic Speech Recognition Engine for US Healthcare.
  - Involves working with Machine Learning Libraries PyTorch and TensorFlow in Python and Kaldi framework for Automatic Speech Recognition (Kaldi, Wav2Vec2, RNN-T, Parakeet, Whisper).
  - Setting up near-realtime ASR Inference endpoints at large scale (Socket-io, Kafka, Websockets, RabbitMQ)
  - Hyper Parameter tuning the models for better performance (Machine Learning, Deep Neural Networks).
  - Handling Raw Audio Data and Processing it to eliminate abnormal data (Sox, Bash, Python, Data Analysis, Pandas, Numpy, Librosa).
  - Developing In-Domain Bencharking Test-set for Rapid Model Evaluation.
  - Automating Data Collection for Active Learning and Strategizing Data Strategy at Organisation Level.

- **Language Modelling [n-gram] for ASR contextual biasing**

  - Cleaning and Pre-processing Text Data at large scale (>10GB) for N-gram Language Model training using SRILM/KenLM toolkit(Pandas, Regex, Bash, Python, GNU-Parallel, Dask, OpenRefine).
  - Designing/Developing/Managing Pipelines with lexicon filtering and Web scraping for Automated LM Generation and Evaluation (Perplexity, Bash, Cron jobs, Selenium, ScraPy).
  - Managing metadata Databases for Data Versioning and Data Lineage (SQL, MongoDB)

EDUCATION

Indian Institute of Technology Madras   **October 2020 - December 2023**
B. S in Data Science and Applications

Sreenidhi Institute of Science and Technology, Hyderabad   **July 2016 - June 2018 (Matriculated)**
B.Tech in Computer Science and Engineering

CERTIFICATIONS

- **Deep Learning Specialization** | T9P329J56AJ8 | Coursera 🔗

- **Machine Learning Engineering for Production (MLOps) Specialization** | 3WEC7TGKXHJN | Coursera 🔗

- **Natural Language Processing Specialization** | JWBBB93TSK7Q | Coursera 🔗

HONORS / ACHIEVEMENTS / VOLUNTEERING / MENTIONS

- **Teaching Assistant | Introduction to Big Data 🔗 | IIT-Madras (Jan 2024 - Present)**

- **Experience as Freelance Flutter and UI/UX Developer (2018-2020)**

- **Technical Head (campus role at ArtsClub, SNIST) (2017-2018)**

- **Volunteered for Global Entrepreneurship Summit, Hyderabad, 2018**

- **Awardee of Undergraduate National Scholarship by Central Government (2016-2018)**