Vatsal Parsaniya

Data Scientist

(+91) 84604-22610
vatsalparsaniya@gmail.com
vatsalparsaniya.com
github.com/Vatsalparsaniya
linkedin.com/in/vatsal-parsaniya

Data Science professional with over 3 years of expertise in Data Science, Information Retrieval, NLP, and microservice development. I have hands-on experience in innovating new products using Data Science and Machine Learning. I have successfully collaborated with cross-functional project teams, SMEs, and led teams with multiple data scientists and engineers. I am passionate about multilingual search systems and extracting valuable insights from complex data.

SKILLS

- **Data Stores :** Elasticsearch, Milvus, Solr, MongoDB, Redis, PostgreSQL
- **Development Tools :** Git, Curl, Jupyter Notebook, PyCharm, Postman
- **Backend Tools :** FastAPI, Airflow, Docker, Azure, Jenkins(CI/CD)
- **Observability :** Newrelic, Loggly, Pyinstrument
- **Statistics / Machine Learning / Deep Learning / NLP :**
  - **Frameworks for NLP :** NLTK, Spacy, PyTorch, Pandas, Scikit-Learn, Text Blob
  - **NLP Models Used :** BERT, RoBERTa, Elastic-ELSER, ALBERT, T5, LLM
  - **Model Deployment & Lifecycle :** NVIDIA Triton, MLflow
  - **ML Algorithms Implemented :** Linear Regression, Logistic Regression, XGBoost, KNN, KMeans, PCA, TSNE, TF-IDF, Word2Vec, Ensemble Algorithms, Topic Modeling
  - **Visualization :** Elastic-Kibana, Metabase, Matplotlib, Seaborn, Plotly
  - **Application Demo :** Gradio, Streamlit

PROFESSIONAL EXPERIENCE

Embibe **Sep 2021 — Present**
Data Scientist ***Bangalore***

- Closely collaborate with the product team within the Discovery Search Science group to transform intricate business challenges into data science problem statements. My role revolves around enhancing the search experience for users by optimizing multilingual search outcomes across a wide range of customer products and internal tools and also engage in both offline and online assessments with the engineering team to ensure the feasibility of our solutions, utilizing methodologies like NLP and Rule-Based systems.
- **Retrieval Augmented Generation (RAG)**:
  - Involved in development of Retrieval Augmented Generation for search and chatbot applications. This entails retrieving academic content with ontologies information and dynamically selecting prompts to generate contextually relevant responses from large language models(LLM).
- **Multilingual Hybrid Search**:
  - Developed search capabilities in 11 Indic languages, including query understanding (QU) and query expansion (QE) modules, enabling users to seamlessly search in any language.
  - Conducted benchmark evaluations on various vector databases, including Milvus, Qdrant, Elasticsearch, and Solr, with a primary focus on retrieval latency and vector index types.
  - Integrated a fine-tuned embedding model into the inference server and incorporated a vector database into the hybrid search pipeline, enabling semantic search capabilities.
  - Established a search feedback pipeline using the Gradio Interface for SME validation and the evaluation of various search algorithms.
  - Setup search utilization dashboard by consuming user event logs to monitor and measure search performance metrics.
  - Achieved a substantial 8% increase in Click-Through Rate (CTR) by optimizing the intent and entity based ranking algorithm, contributing significantly to the system's overall performance.
- **Entity Extraction (NER)**:
  - Designed and implemented a scalable text entity extraction algorithm integrated into a microservice. This system identifies academic entities through the utilization of multiple ontology datasets, synonym dictionaries, and a spellcheck mechanism, all seamlessly integrated with the Solr analyzer to achieve heightened accuracy.
  - The algorithm serves as a core service in various client products, tasked with highlighting academic entities and retrieving associated academic content.
- **Spell Check**:
  - Spell corrector is fundamental component of our NLU pipeline and search engine. Worked on building a corpus for spellchecking using user search query data. I implemented a hybrid spellcheck algorithm that combines popularity-based and context-based approaches. This algorithm successfully improved the accuracy of our system, resulting in the correction of approximately 37% of total user search terms with an impressive accuracy of around 79%.

Intellica.ai **Nov 2020 — Sep 2021**
Machine Learning Engineer (Intern + Full Time) ***Ahmedabad***

- Collaborated on the development of a real-time telephonic conversational AI system aimed at effectively streamlining the interview pre-screening process.
- Enhanced the speech-to-text pipeline for Indian English accents through the application of transfer learning on a deep-speech (STT) model.
- Designed a microservice that includes an evaluation system for the Montreal Cognitive Assessment (MoCA), integrating both computer vision-based drawing evaluation and context-based answer evaluation.

EDUCATION

- **Pandit Deendayal Energy University (PDEU), Gandhinager** 2017 - 2021
  - Bachelor of Engineering (B.E.) in **Information and Communication Technology** with CGPA: 9.2/10.

ACHIEVEMENTS

- 1st Runner up & Best Pitch Award in
- Kaggle 3X Expert
  - Classifies drugs based on their biological activity, Mechanisms of Action (MOA), 208/4373 (Silver Medal).
  - Multi-label classification, SIIM-ISIC Melanoma Classification : 127/3314 (Silver Medal)

COURSES

- Generative AI with Large Language Models, Coursera
- DeepLearning Specialization, Coursera
- Coursera Machine Learning, Coursera
- Time Series with Python (SKILL TRACK), Data Camp