Kshitij Patel

Los Angeles, CA • patelksh@usc.edu • 774-262-8275 • linkedin.com/in/kshitij-patel-20 • github.com/kapatel20

Education

Master of Science, Computer Science **Aug 2022-May 2024**
University of Southern California, Los Angeles, California **GPA: 3.83**
Course Producer: DSCI 552 : Machine Learning for Data Science

Bachelor of Technology, Computer Engineering **Aug 2018-May 2022**
Pandit Deendayal Energy University, Gandhinagar, India **GPA: 4.00**

Technical Skills

Programming: Python, R, SQL, C++, C#(.Net Framework), C, Java, JavaScript ES6, PHP
Data Science/ML: Statistical Analysis, ETL Pipelines, Regression, Classification, Clustering, Anomaly Detection, Recommendation System, Sentiment Analysis, TensorFlow, PyTorch, Transformers, NLTK, PySpark, Scikit-learn, Pandas
Tools/Database: GIT, Tableau, AWS(S3, Kinesis), GCP, Anaconda Environment, Elastic Search, Kibana, Hadoop, MapReduce, Spark

Experience

Software Development Intern, *Ria Money Transfer*, Buena Park, USA **Jun 2023-Aug 2023**
- Leveraged elastic-search and kibana to perform analysis of log files. Preprocessed time-series data and trained LSTM-based model to forecast trends of API failures. Achieved MSE of 4.287 and contributed to enhance decision-making capabilities.
- Demonstrated adeptness in Large Language Models(LLMs) and HuggingFace Transformers library to fine-tune GPT-2, RoBERTa and ALBERT models to develop in-house question-answering system for company's API documentations.

Machine Learning Intern, *Silver Touch Technologies Ltd.*, Ahmedabad, India **Jan 2022-Apr 2022**
- Employed YouTube APIs to extract meta-data about 40000 videos and formulated advanced content recommendation engines utilizing BERT, FastText, and TF-IDF Vectorizer to provide personalized suggestions, yielding Intra-Similarity score of 0.7233.
- Productionalized recommendation system using Flask-based REST API for deployment and maintained dataset.

Undergraduate Research Assistant, *Pandit Deendayal Energy University*, Gandhinagar, India **Aug 2021-Dec 2021**
- Conducted research and leveraged transfer learning and neural networks on MURA dataset for fracture detection in 7 body parts.
- Applied Edge Detection and Image Sharpening Techniques to enhance base model accuracy by 5%. Achieved average Precision value of 84.55% and outperformed Stanford ML Group's performance in case of Fingers by 11%.

Data science Intern, *Silver Touch Technologies Ltd.*, Ahmedabad, India **May 2021-Sep 2021**
- Engineered ETL pipeline to perform exploratory data analysis of about 20000+ schools. Adapted Pandas for data cleaning and preprocessing and applied statistical analysis and visualization libraries to translate data insights into graphical representations.

Projects

Breast Cancer Classification **Dec 2023-Jan 2024**
Python, Pandas, Scikit-Learn, Data Visualization, Machine Learning, Classification
- Implemented Monte-Carlo Simulation and evaluated 4 different classifiers including Logistic Regression, Random Forest, XGBoost, and Support Vector Classifier for supervised learning and self-training for semi-supervised learning over 30 iterations.
- Analyzed and compared results with K-Means and Spectral Clustering with majority pooling for unsupervised classification.

Large Scale Inverted Index using MapReduce **Oct 2023-Dec 2023**
Python, Hadoop, PySpark, MapReduce, Pandas, Regex
- Analyzed multiple text files containing 500000+ words to generate unigram and bigram inverted indexes using PySpark.

Generating SQL Queries from Natural Language **Jan 2023-May 2023**
Python, PyTorch, Natural Language Processing, Semantic Parsing
- Implemented 5 BiLSTM Seq2Vec models in PyTorch based on ACL publication for tasks including selecting aggregation functions, selecting operators, and selecting columns in SELECT and WHERE clauses.
- Integrated attention mechanism between columns and questions to enhance the baseline accuracy by more than 10%.

Time Series Classification for Human Activity Recognition **Jan 2023-May 2023**
Python, Scikit-Learn, Statsmodels, Statistical Analysis, Time-Series Analysis
- Conducted time-series analysis for seven different human activities, extracting seven time-domain features for each activity.
- Utilized cross-validation and backward selection methods to find statistically significant features and trained classification models to classify activity, achieving an accuracy of 94.73% using XGBoost classifier.

Sentiment Analysis of Amazon Customer Review Dataset **Aug 2022-Dec 2022**
Python, PyTorch, Scikit-Learn, NLTK, Gensim, NLP, Sentiment Analysis, Machine Learning, RNN, LSTM, GRU
- Preprocessed data, extracted TF-IDF word embeddings and experimented models including SVM, RNN, LSTM, and GRU to execute multiclass classification. Utilized word2vec-google-news-300 for word embeddings to increase base accuracy by 5%.