Soham Pachpande

+1 (858) 319-5365 | soham.pachpande@gmail.com | linkedin.com/in/sohampachpande | sohampachpande.github.io

Education

University of California San Diego Sep 2021 – Present
MS in Computer Science, GPA: 4/4 *San Diego, CA*

Indian Institute of Technology Gandhinagar Jul 2016 – Dec 2020
B.Tech in Computer Science and Engineering, GPA: 8.35/10 *Gandhinagar, India*

Technical Skills

Programming: Python, Java
Tools: Git, Google Cloud Platform, Linux
Libraries & Frameworks: PyTorch, Pandas, Numpy, OpenCV, Apache Beam, PySpark, Scikit-learn

Relevant Experiences

HSBC Technology Sep 2020 – Aug 2021
Software Engineer, Payments Data Platform Team *Remote*

- Developed streaming ETL pipelines in Apache Beam to ingest payment transaction messages using PubSub, enrich the messages using customer preference data and publish messages to downstream systems to build HSBC's cloud-first tools that process high volume payments transactions and credit/debit alerts
- Introduced usage of BigTable to store the customer preferences data and reduced the pipeline latency by 60%
- Worked with consumer teams to understand the data governance and archival requirements and developed PySpark batch data pipeline to ingest and archive 1 Million+ SWIFT XML payment messages daily
- Provided support for Hadoop based Hive production systems and optimized their migration to Big Query by analyzing data types, cleaning redundant fields, and table partitioning

Mahindra Group May 2019 – Jul 2019
Data Science Intern *Mumbai, India*

- Designed a deep neural network model in PyTorch using the U-Net architecture to perform image segmentation on Sentinel-2 satellite imagery and find potential unused land for affordable housing projects across India
- Took initiative to collaborate with Sales team and developed an interactive map tool in Python to retrieve, analyze, rank and visualize social infrastructure at any given zip code to help customers choose their ideal house

Projects

Image Classification using Hyperdimensional Computing | SEELab, UC San Diego Oct - Dec 2021

- Designed a training framework to train similarity based Hyperdimensional classifier and shallow Convolutional Neural Network(CNN) in a combined manner to achieve a 4% accuracy improvement over the CNN on the 10-class Fashion MNIST classification task

Data Deduplication using Machine Learning Oct - Nov 2021

- Designed and trained machine learning(ML) methods to streamline data preparation by identifying and removing categorical duplicates in tabular datasets with an accuracy of 96% and improve performance of downstream ML tasks
- Performed data preparation to generate duplicate and non-duplicate word pairs for training, validation and test data from a collection of 33 tabular datasets

NLPExplorer | Webapp: *nlpexplorer.org* | Published at ***ECIR 2020*** Aug - Nov 2019

- Built an online search engine and web application (*4000+ monthly users post publication*) to store, analyse and visualise Natural Language Processing(NLP) Literature with an aim to make research more accessible
- Developed a system in Shell and Python to periodically mine research article data from ACL Anthology, apply OCR, and extract paper topics, similarities, and citation graphs using NLP techniques on extracted textual data
- Processed and stored data of 64,520+ papers and 723,976+ citations in MongoDB and Elasticsearch databases

Water Conservation @ IIT Gandhinagar | Published at ***DATA 2020*** Jan - Jun 2019

- Interfaced *66* sensors and engineered a API based system in Python to collect over 190MB data daily to track water consumption, solar energy production and user occupancy at IIT Gandhinagar