Soham Pachpande

858-319-5365 | spachpan@ucsd.edu | linkedin.com/in/sohampachpande | github.com/sohampachpande

Education

University of California, San Diego Sep 2021 – Present
MS in Computer Science *Ongoing*

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

Experience

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

- Developed ETL Streaming Data Pipelines (in Apache Beam/Java on Google Cloud Platform) to generate Payment Transaction Alerts and perform State Management logic to process Payments. Automated an Error Logging framework to track, categorize, and broadcast the error logs in data pipelines to aid system manageability
- Developed batch data pipelines using PySpark on Cloud VMs to process and archive 1 Million+ XML payment messages daily to support SWIFT transactions at HSBC

Mahindra Group May 2019 – Aug 2019
Data Science Intern

- Developed an end to end Image Segmentation model (in PyTorch deployed on Azure Cloud platform) to identify vacant land parcels from satellite imagery. Conducted model selection and hyper parameter tuning experiments to compare U-Net and Mask R-CNN neural network architectures
- Designed javascript map visualization tools to retrieve, analyze and report real estate pricing trends and social infrastructure for business intelligence and marketing teams at Mahindra Group

Mojo Networks May 2018 – Jul 2018
Intern, Member of Technical Staff

- Developed a collaborative algorithm to track the location of consumer WiFi devices using existing infrastructure in linear time complexity and achieves less than 2.5m mean localization error. Key contribution was utilizing interaction between 2 access points to calibrate the algorithm for mitigating environmental noise

Projects

Hierarchical Image Classification using Hyper-dimensional Computing(HDC) Ongoing Research
SEELab, UC San Diego

- Developing image classification algorithms that extract features from images using SIFT, color histogram, and CNN, encode features to Hyper-dimensional(HD) randomized data representations and perform inference using HD vector space operations such as bundling and binding

Data Deduplication in Dirty Tabular Data

- Performed data preparation, feature engineering (using character-level n-grams, Similarity Metrics and Vector Embedding), and applied Machine Learning models (Logistic Regression, Random Forest and ensemble models) to classify whether two categorical data points are duplicates with an accuracy of *98%*

Publications

NLPExplorer : Exploring the Universe of NLP Papers | Webapp: *nlpexplorer.org* | ***ECIR 2020***

- Developed a system to periodically mine research paper PDFs from ACL Anthology, apply OCR, index papers, visualize and designed a web app (*4000+ monthly users post publication*) to make NLP research more accessible. Derived statistics such as topic modeling, citation graph network, and paper similarity from the data

Lessons from Large Scale Campus Deployment | ***DATA 2020***

- Engineered a system to track water consumption, electricity, solar produce and user occupancy using *66* sensors and existing WiFi infrastructure to collect $\sim 190MB$ data daily with an aim to reduce water and electricity wastage

Technical Skills

Languages/Tools	Python, Java, LaTeX, Git
Data Stack	SQL, Apache Beam, PySpark Google Cloud Platform
Frameworks/Libraries	PyTorch, Flask, NumPy, Pandas, Scikit-learn