Soham Pachpande

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

Education

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

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

Technical Skills

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

Experience

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

- Developed Streaming Data pipelines using Apache Beam, Big Table and Google Cloud Platform to process 450 messages per sec. My work supported Payment State Management and Transaction Alert systems and accelerated cloud adaptation of HSBC's data platform
- Developed batch data pipelines using PySpark to process and archive 1 Million+ XML payment messages daily to support SWIFT transactions
- Provided support for production systems and contributed towards designing an Error Logging Framework to track and record errors in streaming data pipelines to aid system manageability

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

- Developed an Image Segmentation model using U-Net architecture to identify empty land parcels within cities from satellite imagery with an accuracy of 85%. Trained and Deployed model using PyTorch and Microsoft Azure

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

- Developed a collaborative algorithm to track the location of consumer WiFi devices using existing infrastructure in linear time complexity that achieves less than 2.5m mean localization error in Python packaged as a RESTful API. Key contribution was utilizing interaction between 2 access points to mitigate environmental noise

Projects

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

- Developed a system using Bash and Python to periodically mine research article metadata and PDF's from ACL Anthology, apply OCR, index papers and derive statistics such as paper topics, citation graphs and similar papers. Stored retrieved data in MongoDB and elasticsearch
- Developed a full-stack web application (*4000+ monthly users post publication*) and RESTful API using Flask to visualise derived statistics and open source our data with an aim to make research more accessible

Data Deduplication in Dirty Tabular Data

- Performed feature engineering, and applied Machine Learning models (Regression, Trees and Neural Networks) to classify whether two categorical data points are duplicates. Achieved a best result of *96%* accuracy using Random Forest Classifier trained on string similarity features on bigrams and trigrams

Agribot | 3^{rd} Best Paper Award at ICSTEM-Vibrant Gujarat 2019

- Built an end to end AI driven chat-bot to answer agricultural queries of Indian Farmers using embeddings trained on govt. of India's farmer hotline call logs. Contributed to data mining, cleaning and feature extraction processes

Publications

- Parmar, M., Jain, N., Jain, P., Sahit, P. J., **Pachpande, S.**, Singh, S., & Singh, M. (2020). **NLPExplorer: Exploring the Universe of NLP papers**. In *Advances in Information Retrieval, ECIR 2020*
- Adhikary, R., **Pachpande, S.**, & Batra, N. (2020). **Lessons from large scale campus deployment**. In *Proceedings of the Third Workshop on Data: Acquisition To Analysis*