Soham Ray

sohamray19.github.io
+1 (917) 753-6259
sr2259@cornell.edu
linkedin.com/in/sohamray19
github.com/sohamray19

EDUCATION

Cornell University, Ithaca, NY | Master's degree in Computer Science
2019 — 2020 ▸ **GPA 3.91/4.0**
Large Scale Machine Learning | Natural Language Processing | Machine Learning for Intelligent Systems | Computational Linguistics

EXPERIENCE

Afiniti | Data Scientist
Aug 2020 — **Present** ▸ Technologies Used: R | Python | SQL
- Analysed changing trends in customer support routing data to boost revenue gain from ~ **0.5%** to ~**5%** (~$500,000)
- Built and maintained statistical models for Afiniti's 4th largest client, using R, SQL and python.

Cornell University | NLP Researcher (Part-time)
Sept 2020 — **Present** ▸ Technologies Used: Python, TensorFlow
- Leading a research team with Prof. Claire Cardie to predict the next tactic in theorem proving using openAI GPT-2
- Building a NLP model to leverage Cornell's NuPRL/arxiv data using tensorflow and Huggingface

Cybage | Machine Learning Engineer Intern
Feb 2019 — May 2019 ▸ Technologies Used: Keras | Tensorflow | NLTK | Scikit-Learn
- Developed software to analyze sentiment and abstracted the product domain so it can be used for any feedback analysis task
- Improved classification accuracy by **11%** on inplace system, and scored an accuracy of **81%** on data from un-trained domains

ASquared IoT | Machine Learning Engineer Intern
July 2018 — May 2019 ▸ Libraries Used: Tensorflow | Scipy | Librosa | Scikit-Learn
- Developed a web application in python to classify welding sound files based on quality with a projected accuracy of 71%
- Constructed denoising models using statistical and ML methods to improve projected classification accuracy to **84%**

Exadatum | Big Data Intern
July 2018 — Dec 2018 ▸ Technologies Used: Spark | Kafka | Hive | Tensorflow | REST API | Flask | PowerBI | Maven | Eclipse
- Developed Xflow, a production-ready real-time streaming ETL data pipeline, with pluggable ML model for sentiment analysis
- Designed data visualization dashboards illustrating various Key Performance Indicators (KPIs) of a Fortune 100 Company

SELECTED PROJECTS

distilBertNQ: Developed a training script for Google's Open Domain Question Answering Task using HuggingFace's distilBert model, low precision GPU computing, and a custom optimizer to understand and answer a question given a raw wikipedia page. F1-score: **59%**

Deceptive Opinion Spam Classification: Constructed an opinion spam classifier to weed out fake reviews written by real people. Explored usage of language models and naive bayes classifier using different linguistic and preprocessing techniques. Accuracy: **91%**

Metaphor Detection with Sequence Labeling Models: Implemented pos tagging and metaphor detection on a given text corpus. Explored feature engineering, hidden markov models and feedforward neural networks to get the observation probability matrix. F1 Score: **65%**

Air Quality Modelling (Master's Project): Designed data pipelines and ML models to predict hotspots hyper-locally and observe spatio-temporal patterns in air pollution in global metropolitan areas to help curb rising air pollution levels worldwide.

quoteBack - Movie Quote Suggester: Constructed a web-app using wit.ai, web scraping, and TensorFlow to read messages on FB messenger and suggest movie quotes as responses based on sentence similarity and intent-bucketing.

TECHNICAL SKILLS

Experience in building scalable web applications/ML models in Java/Python with strong competencies in data structures, algorithms, object-oriented design, data engineering, data ETL, testing and scrum; supplemented with expertise in machine learning and NLP.
- **Languages**: Python, R, Java, C++, Scala, Solidity, SQL, HTML, CSS
- **Frameworks**: **Machine Learning** (Tensorflow, Pytorch, NLP frameworks, Scikit), **Data** (Apache Kafka, Spark, Flume, Avro, Hive, AWS), **Web Dev** (Maven, Git, Java EE, Flask, RESTful services, Eclipse), **Blockchain** (Hyperledger, Ethereum)

CERTIFICATIONS

Coursera ▸ DeepLearning.ai Specialization

Udemy ▸ Data Science for Professionals