Soham Ray

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

EDUCATION

CORNELL UNIVERSITY, ITHACA, NY | MASTER of ENGINEERING IN COMPUTER SCIENCE

Graduated **May 2020** ▸ **GPA 3.9/4.0**

Large Scale Machine Learning | Natural Language Processing | Machine Learning for Intelligent Systems | Computational Linguistics

Master's Project: Designed data engineering pipelines and ML models to predict hotspots hyper-locally and observe spatio-temporal patterns in air pollution in global metropolitan areas (London and Oakland) to help curb rising air pollution levels worldwide. RMSE: 0.06

UNIVERSITY OF PUNE | BACHELOR of ENGINEERING IN INFORMATION TECHNOLOGY

2015 — 2019 ▸ **GPA 8.51/10.0**

Analysis of Algorithms | Data Structures | Cloud Computing | Database Management | Computer Networks | OS

Undergrad Project: Insurance policy database management web-app with role-based access control, interactive payment reminders, payback estimate on insurance maturity, analysis reports for Customers/Agents to ensure a one-stop solution.

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**: **Java, Python**, C++, Scala, Solidity, SQL, HTML, CSS
- **Frameworks**: **Data ETL** (Apache Kafka, Spark, Flume, Avro, Hive, AWS), **Web Dev** (Maven, Git, Java EE, Flask, RESTful services, Eclipse), **Machine Learning** (Tensorflow, Keras, NLP frameworks, Scikit), **Blockchain** (Hyperledger, Ethereum)

PROJECTS AT CORNELL UNIVERSITY

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%**

Story Completion Prediction (Cloze Test): Worked with feature selection and engineering, classification models and pretrained language models (OpenAI GPT-2) to correctly predict the correct ending of a 4 story sentence given 2 endings. Accuracy: **87%**

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.

EXPERIENCE

CYBAGE | SOFTWARE ENGINEERING 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 **80%** on data from un-trained domains

ASQUARED IoT | SOFTWARE ENGINEERING 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 | SOFTWARE ENGINEERING 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

CERTIFICATIONS

Coursera ▸ DeepLearning.ai Specialization

Udemy ▸ Data Science for Professionals