JUDHAJIT ROY

jroy13n96@gmail.com | (872)-218-5124 | **www.linkedin.com/in/judhajit-roy** | **github.com/Judhajit-Roy | judhajit-roy.github.io**

SUMMARY

Software Developer with 3+ years' experience in Big Data development for driving business solutions and in Machine learning and Deep Learning for creating models based on numerical data, images and text primarily using Python, SQL and Spark.

EDUCATION

Master of Science - Computer Science, 4 GPA *Aug 2021 - Present*
University of Illinois at Chicago, USA
Relevant Courses: *Machine Learning, Visualization, Data Structures & Algorithms, Databases, Big Data, Deep Learning, Cloud Computing, Natural Language Processing, Causal Inference and Deep Generative Models*

Bachelor in Electronics & Telecommunications Engineering *Aug 2014 - June 2018*
University of Mumbai, INDIA

TECHNICAL SKILLS

Languages & Tools: SQL, Python, Java, JavaScript, Unix, HTML, AWS, EC2, Lambda, S3, Hadoop, MongoDB, Docker, Agile/Scrum
Frameworks: REST APIs, AngularJS, NodeJS, Spark, PySpark, Hive, Kafka, Presto, Airflow, SparkML, React, gRPC
Machine Learning: Scikit-learn, NLTK, NumPy, Pandas, Seaborn, NLP, CNN, PyTorch, Tensorflow, LSTM, BERT, VAEs, Tableau, D3

WORK EXPERIENCE

Machine Learning Intern – Encyclopedia Britannica | Chicago, IL, USA *May 2022 – Aug 2022*

- Experimented and developed several techniques for named entity recognition and key-phrase extraction from 10k+ articles
- Deployed a pipeline on AWS to automate gathering and transforming XML data into natural language text and extract important key-phrases from top sentences by ranking using a combination of both supervised and unsupervised algorithms

Data Science Intern – Trippers | Monaco *March 2021 - Aug 2021*

A start -up working on perfecting an app to generate travel itineraries and connect travellers.

- Developed efficient algorithms for the core functionality of generating itineraries with the shortest and most preferrable route
- Led proof of concept and collaborated closely with the CEO to build a light-weight recommender system pipeline that recommends cities to visit in Europe based on user preferences and liking by fetching and parsing data from NoSQL Database
- Implemented the recommender system using text similarity for marketing team and reduced manual efforts by 100%

Data Engineer – Reliance Jio Platforms Ltd | Mumbai, India *Aug 2018 - May 2021*

India's Largest LTE Network Provider with Annual Revenue of USD 1.7 Bn.

- Built ETL pipelines on Spark to migrate and aggregate Big Data and cut down the run time from hours to under 15 minutes
- Lead a team of 3 and worked closely with Business team to understand requirements and translate them to technical solutions
- Developed a distributed Sentiment analysis pipeline that classifies customer reviews using SparkML with 86% accuracy
- Scheduled jobs for batch processing, monitored the operation of over 100 ETL jobs and resolved bugs to ensure functionality
- Performed feature generation using Advanced SQL and PySpark operations to curate data and created visualizations on Tableau
- Implemented and maintained operation of real-time streaming of data gathered from OTT Services using Kafka

PROJECTS

Multi-label classification of Research Articles Comparative Study of Machine Learning and Deep Learning Models:

- Classified 20k+ research articles into multiple classes using Natural Language Processing for a comparative study of various models
- Compared the results of the models to determine the best model (Bi-LSTM) and established reasoning for the results obtained

Old Photo Restoration using state of the art Denoising Probabilistic Diffusion Generative Model:

- Repurposed and developed a Diffusion model that restores old and degraded images and generates high quality version images
- Augmented photos with blur and film grain noise to train the model and obtained very clear versions of the noisy images

Log Statistics generation using Map-Reduce Framework on AWS Elastic Map-Reduce Cluster:

- Built end to end pipeline that reads Log Data, processes, and aggregates it using the map-reduce tasks to generate Log Statistics
- Deployed pipeline on AWS, where Map-Reduce tasks running on EMR cluster fetch log data and store results in S3 bucket

Sentiment Analysis on Hotel Reviews:

- Designed and implemented an information retrieval and classification model for sentiment analysis on TripAdvisor reviews
- Crawled 20k+ reviews from TripAdvisor API, and extracted content of JSON responses using Requests module
- Cleaned, parsed and segmented reviews and applied cross-validation technique on the best models to improve the performance