Samar Srivastava

91-70600 04225 • Github • samarsrivastava44@gmail.com • LinkedIn • Visual Github

Education

Dr. APJ Abdul Kalam Technical University • Lucknow, Uttar Pradesh August 2015 – May 2019
Bachelor of Technology • *Computer Science* • Percentage: 71.1

Work Experience

Associate Data Scientist – Avance Consulting Services Pvt. Ltd. Nov 2020 –
Hyderabad, Telangana

- Working on a wide spectrum of technologies in NLP and Data Engineering
- Using Elastic stack to perform ETL to build knowledge base for NLP pipelines.
- Implemented Job Description Parser to extract job based attributes like Job Titles, Locations, Salary, Job Type, Company Name from documents.
- Responsible for managing millions of records spread over multiple ES indices.
- Major tech stack - python3, huggingface tokenizers, nltk, sklearn, Docker, Redis, ElasticSearch, Kibana.
- Using Docker as primary containerization tool.

Machine Learning Engineer – Scanta Inc. April 2019 – July 2020
Gurugram, Haryana

- Worked on data dashboard generation by analysing virtual assistants requests and response to detect anomaly in conversations and report malicious events.
- Setup the pipeline for basic NLP preprocessing like text cleaning, tokenization, generating bag of words, evaluating n-grams. Evaluate similarity between text using cosine similarity, and much more.
- Used Kafka nodes as message brokers, capsulating modules into docker containers.
- Used T-SQL Server as primary database and redis as in secondary database for super-fast data fetching.
- Experimented with Uber AI's Plug Play Language Model to induce personalities in text.
- Responsible for development of a paraphrasing tool using Transformers.
- Deployed 3 products on AWS using various services like EC2, API gateway, and AWS Lambda.
- Responsible for end to end engineering on NLP products pipelines from data mining, data cleaning , to modelling and deployment on cloud.
- Major tech stack - python3, huggingface tokenizers, nltk, sklearn, Docker, T-SQL, HTML, CSS, JS, spacy.

Selected Projects

Text Data Preprocessing Pipleline Scanta Inc 2020
Natural Language Processing

A pipeline to automate tasks like tokenization, lemmatization, and various other NLP preprocessing tasks by setting up a flexible sequential system for performing the above mentioned tasks that earlier need to be done manually one after another, based on requirements.

Predicting Stack Overflow Tags Suggest the tags based on the content in the questions posted on Stackoverflow. 2019
Personal Project

Used multiple classification approaches to determine best predictor.

One-Versus-Rest approach using Logistic Regression with l2 regularizer.

SGDClassifier with One-Versus-Rest approach.

Employee Attrition Rate Prediction ML Competition 2020
Personal Project

Machine learning based approach to predict the attrition rate of employees of an organization to help management in keeping them.

Classification of Business Licence Status Goal is to perform multi-class classification of the business license status 2019
Challenger Project

Relies extensively on the quality of feature engineering.

Performed extensive data transformation, feature generation, and feature importance analysis.

Performed re-sampling for highly imbalanced data.

Used XGBoost as final classifier achieving score percentile of 76(f1-score).

Technical Skills

- Programming languages: Python, T-SQL, C, sqlite3
- Other framework experience:Numpy, Pandas, Flask, Scikit-learn, Keras, nltk, spacy, tokenizers
- Deployment experience: Docker, pypi, EC2, Lightsail, API-Gateway, AWS Lambda