Samar Srivastava

+91-945-7068-769 • 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

Central Board Of Secondary Education • Delhi — March 2014 – March 2015
Higher Secondary School • *Physics, Chemistry & Mathematics Major*

Work Experience

Machine Learning Engineer – Scanta Inc. — April 2019 – Present
Gurugram, Haryana

- Working on analysing virtual assistants requests and response to detect anomaly in conversations.
- 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.
- Calculating tf-idf values over conversations, performing stemming, lemmatization.
- Working closely on production level deployment.
- Used Kafka nodes as message brokers, capsulating modules into docker containers.
- Used SQL Server as in memory database for superfast 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.
- Packages used - nltk, scikit-learn, stanford NLP, and more.

Selected Projects

Covid 19 India EDA Kernel Kaggle — 2020
Personal Project

A notebook dedicated to data visualization and geospatial analysis of COVID19 Pandemic in India.The notebook visualizes the current location of COVID19 patients in India, and the effects of COVID19 pandemic in India to help understand the effect of the outbreak demographically. The Exploratory Data Analysis showed some really interesting insights from the data.

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 that was there in the question 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-Versu-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, SQL, C, sqlite3
- Other framework experience:Numpy, Pandas, Flask, Seaborn, Scikit-learn, Keras, nltk, spacy, tokenizers
- IaaS: AWS, Azure
- Deployment experience: Docker, Redis, EC2, Lightsail, API-Gateway, AWS Lambda