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*

Technical Skills

Programming Languages • Tools
Python,Elastic Search, Kibana, Apache AirFlow, OOP, Bash, Git, GitHub, YAML

Packages • Frameoworks
sklearn, pandas, numpy, transformers, tokenizers, folium, seaborn, nltk, spacy,Flask, FastAPI, Selenium, BeautifulSoup

Cloud • DevOps
AWS(s3,EC2,Lambda, API Gateway, Cloud9, SageMaker), Docker, Azure

Domain Expertise • Domain Knowledge
NLP, Machine Learning, Data Scraping, Text mining, Text analytics, data cleaning, data pre-processing, geospatial data analysis, data visualizations, Statistics, Tech recruiting

Work Experience

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

- Responsible for developing job description parser (JD Parser) using nltk to scan through large amount of raw documents (emails) and extract information pertaining to job title, company name, compensations, locations, industry. JD Parser helped end users by bringing down job creation process from a couple of minutes to a couple of seconds.
- Built multi platform, reusable, well-tested Boolean Query String Generator (BQSG) using Elastic Search, nltk, & spacy. BQSG is able to generate comprehensive boolean search queries just by taking in as input either of the following : Job Title, Job Description.
- Developed industry classification model using pretrained BERT via transformers pipeline, & deployed as a fastAPI containerized via Docker. Response time on seqeunces with count of upto 2000 tokens was 81 milliseconds.
- Performed Data scraping to collect data into Elasticsearch indices for educational institutes, professional degrees, and then performed pre-processing to build knowledge base for data products.
- Developed Airflow DAGs to perform repetitive task like index cleaning, transportation pipelines to move data from ES to redis This ended up boosting teams productivity by a significant factor.
- Responsible for end to end NLP pipeline starting from text pre-processing to packaging NLP tools for usage in production applications.
- **Tech Stack -** python, fastAPI, elasticsearch, kibana, nltk, spacy, transformers, regex, airflow

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

Personal Projects (available on github)

Employee Attrition Rate Prediction Kaggle 2020
Source Code | Tech Stack - python, data science, nltk, pandas, regression, ML

Classification of Business Licence Status Kaggle 2019
Source Code | Tech Stack - python, data science, nltk, pandas, regression, ML

Predicting Stack Overflow Tags Kaggle 2019
Source Code| Tech Stack - python, data science, nltk, pandas, regression, ML

Movie Genre Classfication Challenger 2019
Source Code| Tech Stack - python, selenium, bs4, data science, nltk, pandas, regression, ML