Samar Srivastava

+91-70600 04225 • Github • samarsrivastava44@gmail.com • LinkedIn • Gurugram, India

Education

Dr. APJ Abdul Kalam Technical University • B.Tech — August 2015 – May 2019
Computer Science

Technical Skills

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

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

Cloud • DevOps
AWS, Docker, Azure

Domain Expertise • Domain Knowledge
NLP, Machine Learning, Data Scraping, Text mining, Text analytics, data cleaning, data pre-processing, data visualizations, Statistics, ChatGPT, Rest APIs

Work Experience

Data Scientist – WiseStep — November 2020 – Present
Gurugram, India

- Lead end to end development of a real-time industry classification pipeline
    - Tags candidate experience to relevant industry
    - Macro F1 Score : 0.67
- Working on text and other unstructured or semi-structured data to extract useful information
    - Parsing of free-form resume to detect and extract candidate name, employers, skills, certifications, employment history;
    - Parsing of job descriptions to detect and extract job title, salary, employer name, assign a standard occupation code;
    - Automated noisy and multi-language data normalization, cleanup, deduplication;
    - Tech stack : Python, NER, Spacy, RegeX, POS Tagging, Kafka, AWS, Flask, Celery
- Developed a lead generation pipeline to source job leads for Recruiting Associates based on keyword search & jobs sourced from Linkedin, other job boards.
- Reduced data duplication of active consultants by identifying & flagging candidates that are being promoted multiple times by different recruiters for a job. This helped in keeping the talent pool unique. The method relies on Levenshtein distance & Jaro-Winkler distance.
- Day to day task involves data collection to Elasticsearch indices, EDA for job titles, company names normalizing them and performing pre-processing to build knowledge base for data products.
- Engineered Airflow DAGs to automate repetitive tasks such as index maintenance and data pipeline orchestration.
- Generate and manage data dashboards for the Customer Success team and stakeholders.
- **Tech Stack -** python, flask, elasticsearch, kibana, nltk, spacy, transformers, regex, airflow, AWS, chatGPT, LLMs

Machine Learning Engineer – Scanta Inc. — April 2019 – October 2020
Gurugram, India

- 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 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)

Classification of Business Licence Status Kaggle — 2020
Source Code | Tech Stack - python, data science, nltk, pandas, regression, ML

Predicting Stack Overflow Tags Kaggle — 2020
Source Code| Tech Stack - python, data science, nltk, pandas, regression, ML