HIMANSHI SINGH

https://github.com/HimanshiSingh004

Data Scientist - AstraZeneca

https://www.hackerrank.com/singh_himanshi01

Contact : +91 8100216695, singh.himanshi004@gmail.com

https://www.linkedin.com/in/himanshi-singh3910a2a1/

Summary

I am associated with AstraZeneca as a Data Scientist in R&D IT. In three plus years of hands-on, the goal has always been to bring Data Science features to production. I have a deep crush in problem solving and natural language processing. Ideation, strategy, applied research, modelling, deployment, scaling and monitoring makes my day.

Experience

Data Scientist at AstraZeneca (Aug'20-Present)

Technologies used (In Production) : Python, Deep Learning(Longformer, Scibert), Machine Learning(Gradboost, Logistic), Ensemble, Active learners, AWS Services(Sagemaker, Cloud Formation, State Machines), Knowledge graph, Triz

- Ideated, got POC approved, Built a team, Developed, Deployed, improved over feedback and horizontally scaled a Summarizer service for Competitive Intelligence Platform
    - Bert based extractive graphical solution which leverages CI ontologies
    - Generates headlines using fine-tuned vanilla t5 model
    - Deployed on AWS, as multi-model endpoint through lambda service on CI platform.
    - Presented pipeline overview at Shaastra'21, IITM Techfest as a guest lecturer
- Developed Safety Surveillance feature which serves our scientists in pharmacovigilance space by filtering out documents of interest for six major products. As an ad-hoc requirement, developed a dummy version to accelerate AZD1222 adverse events tracking
    - An ensemble model with Longformer and NER(Termite by SciBite) featured based Gradboost
    - FDA approved and GxP qualified project, extensive audit trials and documentations
    - Deployed on AWS, with free of manual interference re-training and inference pipelines
    - For phase-2 launch, leveraging active learners in ensemble stage (In progress)

Data Scientist at Xlpat Labs (Sept'19 – Aug'20)

Technologies used : Python, Deep learning (PyTorch, Tensorflow, Transformers, Universal sentence encoder, BERT, GPT-2), Machine Learning(NLTK, xgboost, random forest, k-means etc), Elastic Search(aggregation, MLT), Custom Search, GCP services

- Developed and deployed a real time in browser patent ranking model using USE and cosine similarity in Python. Converted trained model into Nodejs using TensorflowJs and deployed to frontend in order to leverage a faster and lighter model.
- Developed a corpus of 3 billion patent keywords and key phrases using word2vec and doc2vec to make customize keyword suggestions from patent text during patent search. Working on a similar product corpus which will leverage many to many mapping between product and assignees.
- Simultaneously working for two clients of Xlpat labs:
    - Indian Patent Office : Initial screening to separate out atomic and defense category patents, proposed approach is rule based filtering followed by multi label classification using fine-tuned BERT on abstract and first claim text embeddings.
    - Chugai Pharma : Build a model for highly unbalanced (75:1) dataset binary classification problem. Leveraged Universal sentence encoder and annoy indexing on mirrored dataset.

Programmer Analyst at Cognizant (Jun'18 – Aug'19)

Technologies used : Python, Dash APIs, Machine Learning visualization, Arima, Plotly, SQL, Postman, Jenkins

- As a part of E*TRADE (US Investment Banking giant) analytics team, developed and unit tested streaming time series dashboards in Dash framework using plotly to visualize seasonality of data.
- Developed 2D and 3D dashboards (honeycomb, candlesticks) to suffix user stories. Unit tested REST based APIs developed in Django before integration to dashboards. Created test pipelines for batch processing in Jenkins.
- Received appreciation for excellent domain knowledge in Investment banking from E*TRADE(client)

Expertise

- Languages and Techniques : Python, Machine learning, Deep Learning
- Frameworks : PyTorch, Tensorflow/Keras, Scikit-learn, Gradio
- Search engine or Databases : Elastic search, Apache solr, My SQL
- API or UI : Django, Flask, Streamlit
- MLOps: Docker, AWS, Git, Jenkins, Redis

Education

- Maulana Abul Kalam Azad University of Technology (2018) - B.Tech. Electrical Engineering (7.8/10)
    - Academic Project : Text Mining (https://github.com/HimanshiSingh004/Text-Mining-Major) Documents from various research domains were processed and clustered to leverage similar documents in a specific cluster. Batched a dataset of 2million research papers for parallel processing.
- All India Senior Secondary School Examination (2013) - St. Xavier's HS School - PCM (91.4%)
- Bihar School Examination Board (2011) - St. Teresa's high School (85.2%)

Looking for

- Opportunities in NLP Ideation and AI Product management
- Flexibility which inclines towards work from anywhere

Personal Information

- DOB -- 04-12-1995
- Permanent Address -- D/O Mr. Ram Pravesh Singh, Sumitra Sadan (HN. 64), Opposite to Fakirana sister's Society, Bettiah- 845438, Bihar