

Shubham Singh

Data Scientist

Chennai, TN 600113

+91 76188 30599

sk_singh18@outlook.com

EXPERIENCE

Okarango Data Technologies, IITM Research Park, Chennai (TN)

Data Scientist – NLP

JANUARY 2019 - PRESENT

- Finetuned transformer and RNN based models for natural language generation, question-answering, text summarization, entity recognition. Implemented end-to-end text cleaning and processing pipelines.
- Delivered machine learning models that are used at scale in production environments to deliver Natural Language Processing (NLP) - based products and services.
- Contributed to the creation of new project proposals including the creation of business use cases.
- Collaborated across diverse groups from clients to junior team members.

Iwiz Education, IITM Research Park, Chennai (TN)

TA Intern

FEBRUARY 2018 - OCTOBER 2018

- Worked on churn prediction of employees for a microfinance company.
- Worked on the development and deployment of OpenEDX - an open-source learning management platform.
- Synthesized content for probability & statistics for providing data science and ML training to technology professionals.

PROJECTS

Generating Question-Answer Pairs from a Given Text Document

FEBRUARY 2019 - DECEMBER 2020

- Implemented a pipelined system that takes a single paragraph as input and produces a set of QA pairs as output. Finetuned *GPT-2* for question generation and *BERT* for question answering task.
- Created flask APIs for serving the system as web app. The pipeline consisted of the following: (i) answer span selection (ii) question generation module (iii) question-answering module (iv) filtering out bad QA pairs.

Inverse Text Normalization

JANUARY 2021 - PRESENT

- Implementing *Automatic Speech Recognition* post-processing for readability that aims to transform the noisy ASR output into a readable text for humans and downstream tasks while maintaining the semantic meaning of the speaker.
- Preparing a dataset synthesis pipeline that takes a grammatically incorrect sentence as input, processes it through TTS to produce audio files simulating errors in human speech, and convert the audio file back to text using an ASR system.

TECHNICAL SKILLS

Languages: Python, C

Tools: NLTK, Spacy, word2vec, BERT, GPT-2, PyTorch, Pandas, Vscode, Jupyter-notebook, Git

EDUCATION

Banaras Hindu University, Varanasi

Master of Computer Applications

AUGUST 2015 - JUNE 2018

Holkar Science College, Indore

Bachelor of Science (CS)

JULY 2012 - JUNE 2015