YASHASVI MANTHA

Data Scientist | Engineer | NLP Researcher

Education

Bachelor of Computer Engineering 2017 - 2021

Gandhi Institute of Technology - Vizag 8.66 CGPA

Recipient of Young Achievers Award for Academic Year 2020-2021

Key - Skills / Frameworks

• Python	• Docker	• Streamlit	• Vector DB	• Redis
• Plotly	• NiFi	• Pytorch	• Elastic Search	• MLFlow
• Pandas	• Spacy	• SQL	• Kubernetees	• Kafka
• SkLearn	• CI/CD	• Neo4J	• Docx Extraction	

Social

+91-9121837667

yashasvimantha@gmail.com

linkedin.com/in/yashasvimantha/

github.com/YashasviMantha

yashasvimantha.com

Work Experience

ISS - Governance (Institutional Shareholder Services), Mumbai — 2022 Jan - Present

ML Engineer - Innovations Lab

- *Developed **ETL pipelines** for various NLP tasks including NER, Entity Linking, Unsupervised Clustering, QA and scaled them for **Big Data**.*
- *Migrated entire platform to private cloud including 40+ Micro-services, 32+ ML Models, 4 Databases: Postgres, Influx, Vector Database, Redis.*
- *Implemented a scalable **Hybrid Search** based ranking with **BERT** based embedder and BM25 to enable high quality textual search.*
- *Built a Text Summarization application with Streamlit using various **Transformer** Based models for News Event (Extractive & Abstractive).*
- *Developed an Entity Linking pipeline using a inhouse **Knowledge Graph** in parallel with DBPedia for linking Entities with Inhouse Identifiers.*
- *Re-evaluated models using task specific evaluation metrics for classification, clustering and tagging*
- *Designed Data Annotation platforms for building domain specific datasets for a clustering task along with setting Annotator guidelines.*
- *Implemented an Information extraction system from **Word Documents (docx)** along with metadata: Control blocks and wrote parsers.*
- *Modified and developed a **custom graph bridge** finding algorithm for pulling connected data points.*
- *Developed **MCP (Model Control Protocol)** for function calling, Natural language integration with Chat APIs and toolsets.*

Deloitte India (Office of the US), Hyderabad — 2021 Jan - May (Intern), July - Dec (FTE)

Intern (4 months) + Full Stack Engineer (Full time)

Keywords: SQL, Stored Procedures, Agile, Backend Development

- *Implemented a Cache Busting mechanism for an internal only application along with developing complex Power BI dashboards.*
- *Fixed numerous bugs, LSIs, and defects that disturbed the workflow of end-users.*
- *Wrote various seed scripts and stored procedures that automated the process of clearing transaction data and populating seed data in tables.*

CFILT (Centre for Indian Language Technologies), IIT Bombay — 2019 May - Aug

NLP Research Intern under Prof. Pushpak Bhattacharyya; Mentored by Dr. Diptesh Kanojia

Keywords: Wordnet, Glove, BERT, Fasttext, Pandas, Python, Phylogenetics

- *Worked on 14 different Indian languages from the IndoWordNet and created Cross-Lingual word embeddings using Muse and FastText.*
- *Surveyed various methods to calculate a 14x14 Distance matrix for these languages including 3 low resource languages.*
- *Utilized edit distance algorithms like Normalized Levenshtein Distance and Angular Cosine to calculated a 14x14 Distance matrix .*
- *Was responsible for cleaning, script conversion and offset conversion to devanagari script for various Indian languages. Also had to handle large wikipedia dumps for FastText, MUSE model training.*
- *Implemented various tree construction algorithms like UPGMA, WPGMA and Neighbour-joining.*
- *Calculated cross-lingual word embeddings between 14 language pairs to perform phylogeny with SkipGram Model for embeddings.*

Research Publications and Presentations

ACM IKDD CODS And COMAD 2020 (India):

• First author of ACM's CoDS COMAD YRS-Track 2020 titled "**Harnessing Deep Cross-lingual Word Embeddings to Infer Accurate Diachronic Phylogenetic Trees**". Presented at ISB (**Indian School of business**), Hyderabad about how the Indian Languages evolved through a tree.

• Poster Presentation under the CoDS COMAD, Young Researcher's Symposium 2020 at 7th ACM IKDD CoDS 25th COMAD about Inferring Accurate Diachronic Phylogenetic Trees.

Open Source Contributions

- Vespa: Yahoo's Distributed Vector Database (6K stars)
- Budget Tracker: A Bank Statement parser categorizer (1 Star)
- Memgraph: A in-memory Graph Database (2.6K Stars)
- H2OWave: Python Dashboarding Framework (4K Stars)