Dheeraj Dhillon

updated May 16, 2024

Email: dheeraj_d@ec.iitr.ac.in **Homepage**: https://djdhillxn.github.io
Phone: (+91) 946-627-6660

Research Interests

Natural Language Processing, Machine Learning, Probability and Statistics

Education

Indian Institute of Technology Roorkee Roorkee, Uttarakhand, India
B.Tech. in Electronics and Communication Engg. Jul 2019 - May 2023
Advisor: Prof. Vinod Pankajakshan

Work Experience

Gartner Inc Gurgaon, Haryana, India
Associate Data Scientist Feb 2024 - Present

- Assesed text documents for readability by evaluating section-wise distribution of syntactic, semantic, coherance features; summarized with measures of moments, shape and dispersion.
- Calculated the Jaccard similarity index on n-gram sets of all document pairs and a corpus-wide cosine similarity matrix on TF-IDF vectors and document embeddings to analyze text uniqueness.
- Integrated readability and uniqueness metrics into a regression model to quantify their impact on client retention, enabling targeted improvements in document design.
- Developed and trained a semi-supervised training pipeline for multi-class classification, expanding the labeled dataset from 700 to 4600 documents. Conducted feature selection on n-grams to identify key drivers influencing document classification to a particular use case.

HiLabs Inc Bangalore, Karnataka, India
Data Scientist Jul 2023 - Feb 2024

- Developed a named entity recognition model using spaCy to extract data header phrases and default value instructions from raw text, facilitating automated data imputation.
- Deployed a multi-class classification pipeline mapping non-standard data headers to two hundred standardized database names using text embeddings integrated with a Random Forest classifier.
- Implemented an information retrieval process to convert unstructured Excel data above headers into image, applying OCR and a LayoutLM to extract scattered key-value pairs.

Microsoft India R&D Hyderabad, Telangana, India
Data and Applied Scientist Intern May 2022 - Jul 2022

- Executed Topic Modelling using LDA on Microsoft Teams conversation texts; optimized coherance scores through grid search to identify five broad themes and their important features.
- Curated and annotated two novel datasets with custom gold keyphrases for conversation texts.
- Deployed the unsupervised KeyBert model for keyphrase extraction; validated outputs using cosine similarity thresholds against a gold standard benchmark.

Research Experience

Undergraduate Researcher iHub DivyaSampark, under the aegis of Govt. of India, IIT Roorkee
Advisor: Prof. Vinod Pankajakshan Jan 2023 - Jun 2023

- Detection of wildlife species involved in human-wildlife conflict.
- Implemented a real-time thermal wildlife detection system on NVIDIA Jetson with a FLIR Boson camera, achieving 10fps and over 95% mAP in 3-class object detection (persons, elephants & other animals).
- Fine-tuned the YOLOv8 architecture on an enlarged novel dataset of more than 80,000 thermal frames. Facilitated annotations using CVAT, refining YOLOv7 pre-detections with a manual review for over 35,000 thermal images.

Select Projects

Estimation of the Warfarin Dose Implemented Multi Arm Linear Bandits including epsilon greedy, Linear UCB, Thompson Sampling.

Handwritten Mathematical Expression Recognition with a Bidirectionally Trained Transformer, as part of Machine Learning Reproducibilty Challenge.

Dry Beans Dataset Performed eda using dimensionality reduction using PCA, t-SNE, autoencoder. Calculated Feature importance using Decision trees, random forests, drop column techniques.

Honors and Scholarships

Bronze Medalist, Inter IIT Tech Meet 10.0, Mar 2022 Kanpur, India
All India Rank 1045, JEE Advanced, Jun 2019 New Delhi, India
All India Rank 1026, JEE Mains, May 2019 New Delhi, India
National Talent Search Examination, May 2017 Gurgaon, Haryana, India

Skills

Programming Languages: Python, C++, Java, SQL
Frameworks: pyspark, sklearn, nltk, spacy, torch, git
Languages: English, Hindi, Spanish