CHAITHRA K . C

315-728-0657 | [ckoppara@syr.edu](mailto:ckoppara@syr.edu) | [www.linkedin.com/in/chaithra-kc](http://www.linkedin.com/in/chaithra-kc)  | <https://github.com/chaithrakc>

# EDUCATION

## SYRACUSE UNIVERSITY, SCHOOL OF INFORMATION STUDIES Syracuse, New York Master of Science Applied Data Science May 2023

Coursework: Introduction to Data Science, Applied Machine Learning, Applied Deep Learning, Natural Language Processing, Business Analytics, and Data Analysis & Decision-Making

**JAIN UNIVERSITY Bangalore, India**

## Bachelor of Engineering Computer Science Jun 2015

Relevant coursework: Principles of Programming Languages, Java & J2EE, Data Structures, Algorithms, Database Management Systems, Data Mining, Business Intelligence, Design Patterns, Unix System Programming, Linux Internals

# TECHNICAL SKILLS

**Key Skills:** Machine Learning, Natural Language Processing, Deep Learning, Data Cleaning, Analysis and Visualization, Hypothesis Testing, Probability and Statistics

**Programming Languages**: Python, R, Java, SQL, PL/SQL, NoSQL

**Python Libraries**: PySpark, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, NLTK, Spacy, Keras, TensorFlow, Gensim, Beautiful Soup, PDFMiner

**Databases**: MySQL, Oracle DB, Mongo DB

**Tools & Utilities:** Databricks, Apache Spark, Tableau, Advanced MS Excel, MS Access, Perforce, Git, Maven

**IDEs:** Google Colab, Jupyter Notebook, PyCharm, R-Studio, IntelliJ

# WORK EXPERIENCE

## Data Science Intern, RSG Media, New York May 2022 - Aug 2022

## Developed a workflow using Databricks and Apache Spark to automatically feed historical data as input to XGBoost ML Model by converting existing SQL Queries to PySpark. This was utilized in predicting future audience

## Entity Resolution between Viacom’s program inventory & Gracenote data, and IMDb & TMDB implemented using NLP pre-processing techniques, TF-IDF word vectorization, and K-Nearest Neighbor ML model

## Built a data pipeline for ingesting 700 million records of metadata from IMDb website to AWS S3 bucket using Python AWS SDK Boto3. After running various data transformations and validations, data was then loaded to Databricks Delta Table

## Created forecast report using Microsoft Excel to compare predictions between different models by aggregating data from several delta tables using SQL and PySpark

## Converted SQL-based ETL pipelines source code to PySpark

## Senior Software Engineer, Envestnet Yodlee, Bangalore, India Mar 2018 - Sep 2020

* Developed data quality tools in Java to validate holding and transaction data veracity and extrapolate missing financial details based on patterns observed with other users
* Modified sequential to a multi-threaded application to process millions of investment accounts and transactions per day
* Implemented Netflix-Ribbon load balancer to achieve fault tolerance and distribution of application traffic
* Implemented POC for determining suitable database for the project among OLTP (Oracle DB), Mongo DB, and Apache Kudu based on query performance
* Designed Mongo DB Collections, documents with suitable indexing, and configured Spring repositories for application querying
* Created REST API using JAX-RS for sharing reconciliation results with other teams

## Software Engineer, Envestnet Yodlee, Bangalore, India Jul 2015 - Feb 2018

* Built web crawlers using Java Selenium APIs to aggregate users' financial data and personal information from banking websites
* Created PDF parser for retrieving data from bank statements using Python PDFMiner text extraction tool
* Implemented regex validator using Java regular expressions for validating PII
* Performed peer code review and mentored a few new joiners

# ACADEMIC PROJECTS

**Sentiment Analysis of COVID News Articles**  **Mar 2022 – May 2022**

* Data cleaning, pre-processing, and exploratory analysis was done on the dataset of 10K news articles that were collected from IEEE data port
* To transform text into numeric vectors, word vectorizations such as the Bag of Words (BOW) model, BERT word embedding, and XLNet embedding are used
* Using Naïve Bayes classifier and BiLSTM neural network, predicted sentiment polarity - Negative, Positive, and Neutral

**News Summarization**  **Mar 2022 – May 2022**

* Generated coherent summaries of 3K news stories obtained through Kaggle
* Model accuracy of Encoder & Decoder using BiLSTM and Keras embedding layer was 46%. summaries generated by pre-trained T5 Transformer were more precise