Sofia Dutta

Curriculum Vitae, Jan 24, 2021

sofia.dutta@newwave.io | https://sofiadutta.github.io

EDUCATION

Masters Professional Studies	Data Science	University of Maryland, Baltimore County, Baltimore, USA	Fall 2020
Bachelor of Technology	Computer Science and Engineering	West Bengal University of Technology, Kolkata, India	Spring 2010

Capstone project title: QABot: A Chatbot for Open Question Answering using Neural Networks

Advisor: Ergun Simsek, Ph.D.

WORK EXPERIENCE

Jan 2021 to Present	Senior Data Scientist @ NewWave Telecom & Technologies Inc., Woodlawn, MD, USA	Working on the Medicaid Data Quality Analytics (MDQA) project, a next generation data science project focused on creating platforms for analyzing and improving the data quality of US healthcare systems. Implementing an end-to-end Data Science workflow from data acquisition, data processing, data integration, model creation, model validation, machine learning prediction to visual representation. Creation of predictive models using machine learning, deep neural networks, and ensemble methods. Automating the collection of data using tools like Google Cloud Fusion to quickly aggregate data from various sources. Pre-processing of data via data cleansing, transformation, formatting to allow proper consumption by downstream analytics systems. Building scalable, distributed, fault tolerant, load balanced systems for analyzing huge quantities of data, typically referred to as Big-Data analytics, using Google Cloud Fusion tool, Google Cloud Storage systems and Apache Spark technologies. Using knowledge of and skills with various cloud-based platforms and technologies like Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SDK, Google Cloud SQL, Google Big Query, and Google Cloud CDN to process, store and analyze patient data and perform data quality analytics over it. Presenting results using visualization platforms like Looker, Jupyter Notebook and Azure Databricks. Resolving business challenges in an efficient and timely manner. Effectively communicating, presenting the most relevant information, and collaborating with team leads and product owners to ensure a positive impact on the US healthcare system and US citizens' healthcare data. Thinking outside the box and researching ways of improving data security and enabling faster data lookups.

May 2020 to Dec 2020	Data Scientist Intern @ NewWave Telecom & Technologies Inc., Woodlawn, MD, USA	Interned on the Medicaid Data Quality Analytics (MDQA) project and applied data visualization skills acquired during graduate career to carry out several visualization tasks for the team. Used Matplotlib, Seaborn to present data quality outcomes from various quality computation metrics. Learned to use the Looker tool from Google as team was heavily using it. Created an end-to-end design and architecture of the data quality analytics platform. Created the database schema design for the MDQA project. Collaborated with product owner and other engineers on building the first version of the data quality platform for the MDQA project. Created mechanisms for generating fake training data using Python programming to test out efficacy of machine learning algorithms used in the project. Trained machine learning models using thousands of rules for quality computation metrics. Worked on improving computation efficiency by several orders of magnitude by implementing data analysis workflows in Google Cloud Platform (GCP) clusters and using Apache Spark for quality metrics computations. Exported results to Looker for creating intuitive dashboards visualizations. Worked with large quantities of US healthcare data from Centers for Medicare & Medicaid. Created clusters in Google Cloud Platform and installed all required Big-Data analytics packages and software. Configured GCP environment to execute Python programs and connected the cloud infrastructure with Looker's dashboards for delivering computed results to be presented to customers.
Sep 2019 to May 2020	Graduate Student Researcher @ Ebiquity Research Group, UMBC, Baltimore, MD, USA	Authored an Ontology for Smart Home Access Control by extending earlier research in Semantic Web. Ref: https://ebiquity.umbc.edu/paper/owl/id/887/Context-Sensitive-Access-Control-in-Smart-Home-Environments Developed an Android app for handling context-sensitive access control in a Smart Home Environment. Ref: https://github.com/sofiadutta/AndroidSmartLights Created YouTube videos for presentation to the National Institute of Standards and Technology. Ref: First demo - https://www.youtube.com/watch?v=rOWJNGHVcHo Second demo - https://www.youtube.com/watch?v=DZq-oZ_Cv1g Published a paper at the IEEE Big Data Security 2020 conference. Dutta, Sofia, et al. “Context sensitive access control in smart home environments.” *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2020. Ref: https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00018

Nov 2010 To Feb 2018	Software Engineer, IT Analyst @ Tata Consultancy Services, Kolkata, WB, India	Worked for seven different projects for clients of TCS. Led the design and development of software requirements that came from the customer. Built API interfaces using PL/SQL stored procedures for daily usage for clients of TCS. Led meetings to capture requirements from multi-national clients like DHL UK, Staples USA, Hyatt USA, Kaiser-Permanente USA. Carried out change based regression analysis and documented software functional specifications for those changes. Prepared test plans and executed system integration tests and user-acceptance tests. Worked on data migration projects and ensured migration activities were completed in a time-sensitive manner thus saving millions of dollars of lost revenue for the clients. Implemented scripts for data migration of over a billion records while adhering to strict time SLA bounds. Completed client data migration from legacy Oracle Apps (11i) to Oracle ERP Suite (R12). From 2013 - 2018, managed continuous integration and continuous deployment in production environments. Additionally, worked as a team lead ensuring engineers joining the team newly were brought up to speed quickly. As a requirement from the clients and due to needs of the projects, completed certifications for Oracle Apps technologies including, Oracle Global Human Resources Cloud 2017 Implementation Essentials, Oracle Talent Management Cloud 2017 Implementation Essentials, Oracle Global Human Resources Cloud 2016 Implementation Essentials, Oracle Advanced PL/SQL Developer Certified Professional, Oracle E-Business Suite 12 Financial Management Implementation Specialist: Oracle Receivables, Oracle PL/SQL Developer Certified Associate and Oracle SQL Developer Certified Associate.

TECHNICAL SKILLS

Programming	Python, Java, SQL, PL/SQL, T-SQL
Machine Learning / Deep Learning / Big Data tools	PyTorch, Sci-kit Learn, Apache Spark, Keras, Tensorflow, Hive, Hadoop, MLlib, Matplotlib, Seaborn library, Looker, LookML
Enterprise tools	Google Cloud Platform, Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SQL, Google Big Query, Amazon Web Services S3, Azure Databricks, Oracle Global Human Resources Cloud, Oracle Talent Management Cloud, Oracle Financial Management
Back-end tools	Oracle Databases, PostgreSQL, MongoDB
IDEs/Dev tools	Jupyter Notebook, Google Colab, PL/SQL Developer, Git
Domain knowledge	Machine Learning, Deep Learning, Big Data Analytics, Semantic Web

DATA SCIENCE RELATED PROJECTS

Aug 2020 – Dec 2020	QABot: A Chatbot for Open Question Answering using Neural Networks Built "QABot", a Chatbot using the sequence-to-sequence Deep Learning model that utilizes the Encoder Decoder Neural Network architecture combined with Attention Mechanism to answer user search queries. Created a model by training a Deep Neural using the PyTorch Deep Learning Framework. Used the Seq2Seq algorithm that trains a Denoising Auto-Encoder over sequences. Used Recurrent Neural Network architecture that are better at dealing with text sequences. Used a randomized algorithm to choose between Teacher Forcing and Auto-Regressive approaches for model training and Auto-Regressive approach for model evaluation. Additionally, used BERT (Bidirectional Encoder Representations from Transformers) for tokenization and combined Transformer and GPT-2 for model fine tuning. Ref: https://sites.google.com/umbc.edu/data606/fall-2020/sofia-dutta
Jan 2020 – May 2020	Image to image translation using CycleGAN Implemented CycleGAN for an image-to-image translation. Trained an unsupervised image translation model via the Generative Adversarial Network (GAN) architecture using unpaired collections of images from two different domains. CycleGAN has previously been demonstrated on a range of applications and I chose to perform object transfiguration with it. Transforming images of horses to zebras and then back from zebras to horses. Ref: https://sofiadutta.github.io/datascience-ipynbs/pytorch/CycleGAN_Img_Translation_PyTorch_Horse2Zebra.html
Sep 2019 – Dec 2019	Big-Data Analytics and Sentiment Classification with Twitter Stream Data In this project, I worked on large real-time streaming data from Twitter. Performed analytics using PySpark and created visualizations. Created a MyClassifier sentiment classifier via Word2Vec model using Spark MLlib. My project goal was to use PySpark, a Big Data tool and perform analytics driven by the 6Vs of Big Data. This helped me because I anticipate that in a career in Big Data, I would eventually face challenges with petabyte-scale data I wanted to be prepared for that. Ref: https://sofiadutta.github.io/datascience-ipynbs/big-data-analytics/Using_MyClassifier_Twitter_Data_Sentiment_Classification_and_Big_Data_Analytics_on_Spark_Dataframe.html
Jan 2019 – May 2019	Comparison of Word2vec and Doc2Vec model driven Sentiment Analysis using SVM, LR, Keras CNN, Bidirectional LSTM with and without pre-trained Word and Document Embeddings In this project, I worked on applying opinion mining or sentiment analysis via word embedding and document embedding models to carry out sentiment classification of user reviews. I performed this classification on two different datasets. The first dataset contained laptop product reviews from Amazon's website and the second dataset consisted of movie reviews from IMDb's website. Ref: https://sofiadutta.github.io/datascience-ipynbs/sentiment-analysis/Sentiment_Analysis_IMDB_Movie_Review.html
Jan 2019 – May 2019	Exploratory data analysis and data characterization of New York City Film Permits

	Data characterization of New York City Film Permits to figure out where my favorite movies being shot at Ref: https://sofiadutta.github.io/datascience-ipynbs/EDA/Data_Analysis_NYC_Film_Permits.html
Jan 2019 – May 2019	Exploratory data analysis and data characterization of New York City Fire Department Data characterization of New York City Fire Department data to find impactful events through a data analytics path Ref: https://sofiadutta.github.io/datascience-ipynbs/EDA/Data_Analysis_NYC_Fire_Department.html
Jan 2019 – May 2019	Exploratory data analysis of Baltimore City Employee Salaries Exploratory data analysis and data characterization of Baltimore City Employee Salaries to study income inequality in Baltimore City Government Ref: https://sofiadutta.github.io/datascience-ipynbs/EDA/Data_Analysis_Baltimore_City_Salaries.html

PUBLICATION(S)

2020

Dutta, Sofia, et al. "Context sensitive access control in smart home environments." *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2020.

AWARDS

TCS Gems for contribution to the organization, Jan 2017, Tata Consultancy Services

Awarded the Champions of ILP (23-Jan-2017) TCS Gems in appreciation of outstanding contribution to the organization, for being an inspiring role model to colleagues and dedication and commitment to excellence.

"Most likely to slay a Dragon", Jun 2015, DHL UK

For persistence and efforts around Finance Data Migration and resolving defects.

Sahara Sukanya Scholarship Award for Academic Excellence, Dec 2005, Sahara India and Holy Child Institute

Awarded a scholarship for academic excellence in high school

LANGUAGES

English	Fluent in all forms of communication
Bengali	Native language, fluent in all forms of communication
Hindi	Fluent in all forms of communication

References available upon request