DARK DATA EXTRACTION AND ANALYSIS FOR DOCUMENTS

Dark data is the data which is acquired by organizations in the normal course of their operations but not used in any manner to derive insights or for decision making and further business logics.However, the same data could be useful to plan product roadmaps, aid business decisions or optimize operations. This is possible today due to modern techniques of machine learning and data analytics.

The term "dark" does not refer to something evil or illegal. Nor is it specifically about security or privacy. Rather, it's about data that's hidden from view, easy to ignore, and hard to access or analyse. It can include information gathered by sensors, telemetic, feedbacks, surveys, financial information, logs, inactive and Old Versions of Relevant documents, documents and great mass of data buried in images, text, tables, figures, PDFs, csv, mp3, ogg, xlsx, png, doc etc.

Our aim is to **extract** dark data using textract from different documents (**Old Versions of Relevant documents, documents**) and sources, structure the data if possible, For unstructured data attach metadata labels that so they can be easier to find for future analysis. Finally, **analyse** it so that it can be used in the future to bring maximum productivity and ability for organizations to meet consumer's demand. Tools to work with dark data include DeepDive, Snorkel, and Dark Vision.

ASHWIN C (312217205011)

HARISSH N (312217205034)

MAYANK SINGH (312217205047)

Student Name

DR. N. BHALAJI

DR. S. KARTHIKA

Guide Name with Signature