**DATA COLLECTION &**

**PRE-PROCESSING**

**GROUP PROJECT**

**GROUP 25**

Venkata Soma Sundaram Pappu – 12110110

Joydeep Vasudeva – 12010029

Phani Ayyalasomayajula – 12110007

**Executive Summary**

The objective of our **Data Collection & Pre-processing** group project is to create an end-to-end data collection and pre-processing pipeline for **“Grammy Awards”**. Our input data comprises of the "Grammy awards" domain seed for which the data collection from various external sources and pre-processing steps are carried out to convert into a structured source. One of the key objectives set in our data collection is to extract the data from various sources – structured and unstructured and ensure the reliability of the data.

We have started with the seed datafile and started enriching the source by extracting new fields from related “Grammy awards” sources across the world wide web. We have enhanced the seed datafile by adding new information rich and reliable attributes and increased the number of attributes to **40**.

We have started searching for our data sources through – Wikipedia artist pages, Grammy awards site, and leveraged Google search to extract the artist related key attributes. From these sources, we have retrieved related data with key attributes to enhance the datafile. For data extraction, we have utilized the following scraping python libraries which enable to access and parse data from the external data sources:

* **Beautifulsoup** – parsing HTML & XML sources
* **Pandas** – Data manipulation & analysis
* **Wikipedia** – Data analysis and parsing from Wikipedia. Get the snippets of URLs from Wikipedia artist pages, build artist URLs, and the related artist attributes.
* **Google search engine** – Retrieve additional relevant attributes through Google search appending the artist names using the “requests” python library.
* **Sweetviz** - High-density visualizations to perform EDA (Exploratory Data Analysis)
* **Tableau** – For merge/joins of the data from multiple sources via primary key

As the data collection was performed from multiple sources, we have created a unified knowledge database by merging the datasets based on the primary key field. The final unified dataset is stored in a CSV format. After the completion of data collection, we started to clean the data before putting it to use for further analysis. In this preliminary step, we classify the data into Missing data & Noisy data. For missing data, our approach is to ignore the entire tuple of data, and remove the records with missing values from the dataset. For noisy data, we have segmented available data into categories based on their characteristics and then processed it accordingly. For transforming the data, we have used the attribute selection from a pre-constructed range and use them accordingly. For data reduction and accuracy, we have only used the attributes that are highly relevant and discarded unnecessary data.

Key challenges which we encountered in our Data collection & pre-processing are:

* Missing data.
* Data inconsistencies.
* Human error leading with different data types.
* Data manipulation.
* Advance programming.

**Domain and Seed sources:**

* We have started with the seed datafile and started enriching the source by extracting new fields from related “Grammy awards”. Total number of relevant attributes **40**. We have extracted rich information across multiple data sources both structured and unstructured to prepare the most cleaned and accurate dataset.

**Data Sources:**

* Data sources crawled: Wikipedia artist pages, Grammy awards site, and Google search to extract the artist related key attributes. From these sources, we have retrieved related data with key attributes to enhance the datafile.

**Data Crawling & Conversion methods:**

* For data extraction, we have utilized the following scraping python libraries which enable to access and parse data from the external sources
  + **Beautifulsoup** – parsing HTML & XML sources
  + **Pandas** – Data manipulation & analysis
  + **Wikipedia** – Data analysis and parsing from Wikipedia. Get the snippets of URLs from Wikipedia artist pages, build artist URLs, and the related artist attributes.
  + **Google search engine** – Retrieve additional relevant attributes through Google search appending the artist names using the “requests” python library.
  + **Sweetviz** - High-density visualizations to perform EDA (Exploratory Data Analysis)
  + **Tableau** – For merge/joins of the data from multiple sources via primary key

**Key Challenges encountered:**

* Missing data either due to manual input or a random behavior.
* Data inconsistencies due to the variations created in the data sets and when multiple values are stored in the same field, which could lead to Data integrity.
* When storing information under different data types, human error could lead to incorrect representations of the data.
* Data manipulation – When working with software for datafiles, sometimes the data is broken during the regular file open/read/write operations.
* Advance programming needed to convert the desired data sources into a structured data.

**Data cleaning/pre-processing methods:**

* Created a unified knowledge database by merging the datasets based on the primary key field.
* Clean the data before putting it to use for further analysis.
* Classify the data into Missing data & Noisy data. For missing data, our approach is to ignore the entire tuple of data, and remove the records with missing values from the dataset. For noisy data, we have segmented available data into categories based on their characteristics and then processed it accordingly.
* For transforming the data, we have used the attribute selection from a pre-constructed range and use them accordingly.
* For data reduction and accuracy, we have only used the attributes that are highly relevant and discarded unnecessary data.

**Any Observations/ Insights and Analysis on the data collected?**

* We have performed an Exploratory Data Analysis (EDA), by using an open-source python library "Sweetviz".

Visualizations/Insights shared below:

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with medium confidence

**Strategy to enhance the data with crowd sourcing methods:**

* The Grammy awards is a past data and does not apply to the crowd sourcing strategy.

**References and Sources used for this Assignment:**

* Wikipedia
* Discogs.com
* Grammy.com
* Python.org
* https://stackoverflow.com