SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 19AD501 Big Data Analytics

III Year / V Semester

Unit 4 – Data Preparation

Topic 1- Data Munging

Data Munging

Data Munging also referred as Data wrangling can be defined as the process of

- cleaning,
- organizing, and
- transforming

raw data into the desired format for analysts to use for prompt decision-making.

- Often, data munging occurs as a precursor to data analytics or data integration.

- High-quality data is essential for sophisticated data operations.

- The munging process typically begins with a large volume of raw data.

- Data scientists will mung the data into shape by removing any errors or inconsistencies.

Data Munging

- Data Scientist then organize the data according to the destination schema, So that it’s ready to use at the endpoint.
- Munging is generally a permanent data transformation process.
- It further includes data aggregation, data visualization, and training statistical models for prediction.
- Data wrangling is one of the most important steps of the data science process.
- The quality of data analysis is only as good as the quality of data itself, so it is very important to maintain data quality.

Data Munging

The modern data munging process now involves six main steps:

1. Discover
2. Structure
3. Clean
4. Enrich
5. Validate
6. Publish

Data Munging

Discover

- First, the data scientist performs a degree of data exploration.
- This is a first glance at the data to establish the most important patterns.
- It also allows the scientist to identify any major structural issues, such as invalid data formats

Structure

- Raw data might not have an appropriate structure for the intended usage.
- The data scientists will organize and normalize the data so that it’s more manageable.
- This also makes it easier to perform the next steps in the munging process.

Data Munging

Clean

- Raw data can contain corrupt, empty, or invalid cells.
- There may also be values that require conversions, such as dates and currencies.
- Part of the cleaning operation is to ensure there’s consistency across all values.
- For instance, the state in a customer's address might appear as Texas, Tex, or TX.
- The cleaning process will standardize this value for every address.

Enrich

- Data enrichment is the process of filling in missing details by referring to other data sources.
- For example, the raw data might contain partial customer addresses.
- Data enrichment lets you fill in all address fields by looking up the missing values elsewhere, such as in the CRM database or a postal records lookup

Data Munging

Validate

- Finally, it’s time to ensure that all data values are logically consistent.
- This means checking things like whether all phone numbers have nine digits, that there are no numbers in name fields, and that all dates are valid calendar dates.
- Data validation also involves some deeper checks, such as ensuring that all values are compatible with the specified data type.

Publish

- When the data munging process is complete, the data science team will push it towards its final destination.
- Often this is a data repository, where it will integrate with data from other sources.
- This will make the munged data permanently available to all consumers

Data Munging

Data munging deals with the following functionalities.

1. **Data exploration**: Visualization of data is made to analyze and understand the data.

2. **Dealing with missing values**: Having Missing values in the data set has been a common issue when dealing with large data set and care must be taken to replace them. It can be replaced either by mean, mode or just labelling them as NaN value.

3. **Reshaping data**: Here the data is either modified from the addressing of pre-existing data or the data is modified and manipulated according to the requirements.

4. **Filtering data**: The unwanted rows and columns are filtered and removed which makes the data into a compressed format.

5. **Others:** After making the raw data into an efficient dataset, it is bought into useful for data visualization, data analyzing, training the model, etc.

Data Munging

Some examples of basic data munging tools are:

- Spreadsheets / Excel Power Query - It is the most basic manual data wrangling tool
- OpenRefine - An automated data cleaning tool that requires programming skills
- Tabula – It is a tool suited for all data types
- Google DataPrep – It is a data service that explores, cleans, and prepares data
- Data wrangler – It is a data cleaning and transforming tool

THANK YOU