Data Preparation Pipeline

It consist of 3 files:

1. Data Load
2. Secondary Inference
3. Data Prepare

Data Load

It contains a class Dataload which have a method DLoad which takes 2 arguments (both are string) as inputs i.e. path and dataFormat where path is the storage location for the data file which is to be loaded, dataFormat can be one among of “csv”, “json”, “xml” and “excel”. DLoad loads the data and returns a dataframe for the input data.

Secondary Inference

It have a method metaSchema which takes two arguments “df” (data type : DataFrame) and path (data type : String), df is the dataframe for which we need to find secondary schema(having secondary type), path is the storage location where the output json file is to be stored. metaSchema outputs a json file to the mentioned storage path which contains 3 keys i.e. columName, primary, secondary where columnName is the name of the column, primary is the primary datatype (as per org.apache.spark.sql.types), secondary is the secondary datatype (can be one among : “DateType”, “email”, “countrycode”,” geoLocation”, “geoCordinates”, “currency”, “integer” and “double” . “DateType can have any date/time format) inference by method metaSchema.

Data Prepare

It is a user made Transformer(machine learning transformer) which have a transform method which takes a dataframe as input and outputs a processed dataframe. In whole process, it remove links, emojis, special characters, unnecessary spaces, stemming using nlp, converting text into vectors. The output dataframe is same as the input dataframe with an additional column “features” which contains vectors of the whole input row.