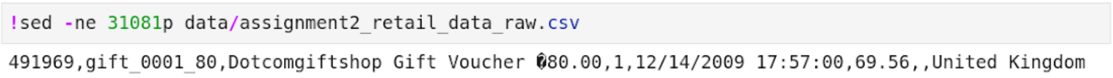
Clarification regarding non-utf-encoded data in retail-dataset file

Vinayak Nayak <2021fc04135@wilp.bits-pilani.ac.in> Thu, Mar 2, 2023 at 9:49 AM
To: "B. SUNIL" <sunilbhutada@wilp.bits-pilani.ac.in>
Bcc: SHREYSI KALRA <2021fc04586@wilp.bits-pilani.ac.in>

Hi Sir,

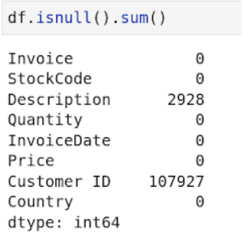
Good Morning!

I am running Hadoop map-reduce using Python on top of Hadoop streaming API. There are non standard i.e. (not `utf-8`) characters also in the provided data. This will cause our map reduce jobs to fail as they wouldn't be rendered in our python map-reduce characters. We will have to convert these into utf-8 values. For eg, consider the following line.



```
!sed -ne 31081p data/assignment2_retail_data_raw.csv
491969,gift_0001_80,Dotcomgiftshop Gift Voucher �80.00,1,12/14/2009 17:57:00,69.56,,United Kingdom
```

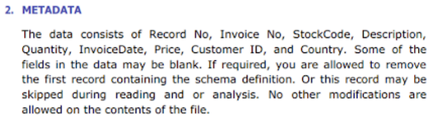
There is a pound sign in the description which is not getting rendered properly, hence we will have to first convert them into `utf-8` characters before passing them to map-reduce engine. Also there are some missing values which are present in the data especially in description and customer ids.



```
df.isnull().sum()

Invoice             0
StockCode           0
Description      2928
Quantity            0
InvoiceDate         0
Price               0
Customer ID    107927
Country             0
dtype: int64
```

Now, one of the requirements says



2. METADATA

The data consists of Record No, Invoice No, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country. Some of the fields in the data may be blank. If required, you are allowed to remove the first record containing the schema definition. Or this record may be skipped during reading and or analysis. No other modifications are allowed on the contents of the file.

Is it fine if we provide justifications for our decisions and process the file so that it could be easily loaded into hive-table or in map reduce jobs for convenience?

Thanks & Regards,
Vinayak Vinod Nayak
2021FC04135
M.Tech (DSE)[2021 October Batch]
Cell: +91-8652380100