Challenge # 3: Profit and Loss statement processing

Problem Class: Transcription-Structured Output

PROBLEM STATEMENT:

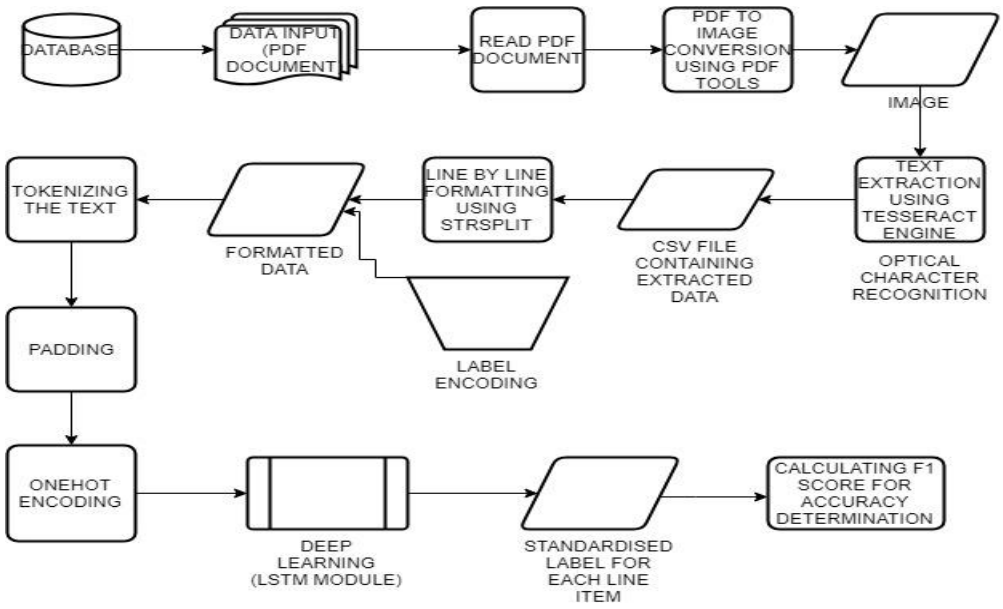
An integral part of any financial statement is Line Items that describe for example nature of earning, expenditure or assets. A line item conveys a particular meaning/financial concept and the same thing can be expressed in different ways. A financial document normally contains more than one line item and each of them conveys a particular meaning/financial concept. The way Line Items are written depends on individuals and no standard is followed. Each financial document will have line items described using different verbiage and certain amount of skill set is needed to understand a financial document and sometimes error arises because of wrong interpretation of the line items.

Therefore there is a pressing need to convert these non-standard line items to a standard format and this can help in avoiding costly error.

PROBLEM DEFINITION:

Conversion of free flowing text description of Line items into a standard description and extraction of a set of corresponding information from an excel document.

SOLUTION:



- The given dataset contains agglomeration of PDF documents containing the financial documents and records of respective companies. The content of each and every document is a scanned image of the particular financial proof (i.e.) the bill or the financial statement, in physical form are scanned as images and then loaded to a PDF file. Each distinguishable PDF document contains the complete financial record of a company.
- As the first step, the documents containing the images of financial records are converted as images (i.e.) every single image from the PDF document is converted to separate distinguishable image in PNG format, since PNG format is a supported format in future processing.
- The PNG image files now containing the financial statements that are converted into texts. OCR (Optical Character Recognition) is utilized for the conversion of images (containing textual financial statements) into text format.
- Once the textual information are extracted, labeling each and every record (row by row) is done. Then based on the labels created, the type of financial statement is differentiated and finally a standardized format of the financial statement is written on an Excel file.

TOOLS UTILIZED:

- RStudio version 3.5.1

  Libraries used (in RStudio)
  - tesseract
  - pdftools
  - lcaret
  - magrittr
  - stats
  - magick
  - text2vec
  - stringi
  - stringr
  - purrrlyr
  - Rcpp
  - XML
  - htmltidy

- tidyverse
- stringi
- xgboost
- magrittr
- rvest
- httr
- keras

- Microsoft Excel to read the CSV file.

STEPS FOR EXECUTION:

1. The given data is in PDF format. Each PDF document contains a number of financial statements as scanned images.
2. The scanned images are first converted as distinguishable image files of PNG format.
3. These images are then fed to the TESSERACT engine for extraction of text from the images. It uses OCR (Optical Chatacter Recognition) for text extraction from images.
4. The extracted data is preprocessed i.e. text alignment, cleaning etc. are done and then labeling is done manually for every line in the extraction file.
5. Then the file is processed in LSTM module and the remaining lables are predicted.
6. Finally the converted financial statements with proper lables (created manually) are written to a CSV file for future reference.
7. For testing the model performance and measure of accuracy, we are calculating F1 score which gives a better accuracy result.

AUTOMATION SCOPE:

- Labeling of line by line extracted data that is stored in a CSV file is done manually. The labels are choosed in reference with valid financial terminologies.

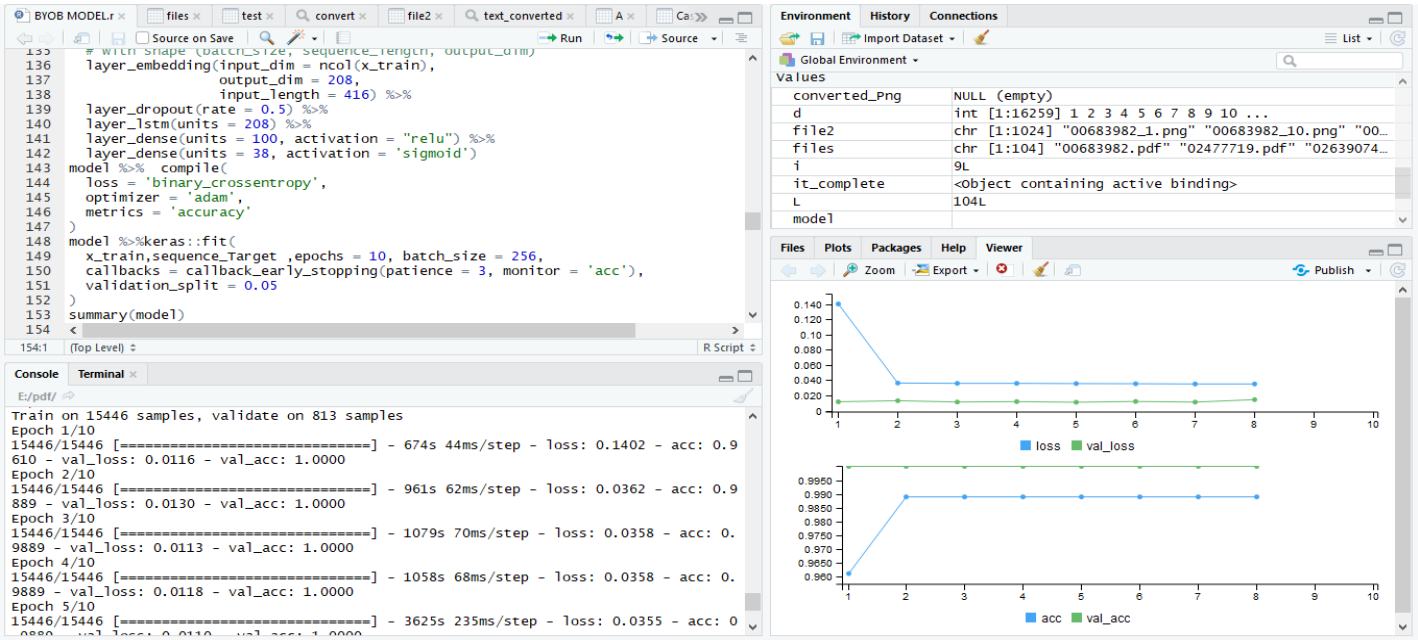
ACCURACY:

- The respective model implented on RStudio gives an accuracy percentage of 98.18%.

MODEL SUMMARY:

```
_____
Layer (type)                 Output Shape              Param
 #
=================================================================
=========
embedding_1 (Embedding)      (None, 416, 208)          86528
_____
_____
dropout_1 (Dropout)          (None, 416, 208)          0
_____
_____
lstm_1 (LSTM)                (None, 208)               34694
4
_____
_____
dense_1 (Dense)              (None, 100)               20900
_____
_____
dense_2 (Dense)              (None, 38)                3838
=================================================================
=========
Total params: 458,210
Trainable params: 458,210
Non-trainable params: 0
_____
```

SCREENSHOT:



LIMITATIONS:

- The conversion from PDF to text could be done better if paid tools were used for building the model.
- As students from Engineering background whatever the labelling of data has been done as to the fullest of our knowledge and capability, as we possess no prior accounting knowledge.