Assignments

Python & Machine Learning

Contents

Assignment 1 – Python Assessment

File

http://www.kepcorp.com/annualreport2018/pdf/keppel-corporation-limited-annual-report-2018.pdf

Page No – 12 of 230

Objective

Extract text blocks from a PDF document with each paragraph as a separate line and text being in reading order going from first column from top to bottom and then into second column and then third column. Dump the output in an excel file.

The solution should be a general solution that can work on different pages of different PDFs.

Resources & Hints

PyMuPDF – PDF Parsing library

Identify columns by looking at the x coordinates difference

OpenCV can be used to combine lines using contours

Submission

Create a github repo, upload the code and the generated output excel along with requirements.txt and share the link.

Assignment 2 – Deep Learning Inferencing

File

http://www.kepcorp.com/annualreport2018/pdf/keppel-corporation-limited-annual-report-2018.pdf

Page No – 69 of 230

Objective

Detect and locate tables inside a page using DNN, then parse the table to extract a dataframe that should not lose any information and structure of the table. The output should be an excel file containing the extracted table.

The solution should be a general solution that can work on different pages of different PDFs.

Resources & Hints

Camelot-py – Table parsing library

Cascadetabnet – A pre-trained DNN for table detection available on github

Submission

Create a github repo, upload the code and the generated output excel along with requirements.txt and share the link.

Assignment 3 – NLP Training

Files

The dataset can be downloaded from here:

https://groups.csail.mit.edu/sls/downloads/restaurant/restauranttrain.bio

Objective

Train a named entity recognition model to identify the various entities present in the database. The final input will be a text string and the program should output all the named entities along with the entity text in dictionary format.

Resources & Hints

Spacy NLP

RASA NLP

Submission

Create a github repo, upload the code along with requirements.txt and share the link. The code should either use argparse to pass text as input or should have a callable function in which the text can be passed and it return the entities,