Challenge #2
OCR De-Noising

The BYOB Automation Challenge 2018

Submission By:

Akshay Sharma

Ashish Rana

Technical Components

Platform

- Windows 10

Tools

- Poppler for Windows
- Python

Language Libraries

- elementtree
- python-docx
- pyspellchecker

Problem Overview

- Majority of the business documents today are in Portable Document Format (PDF); a file format for capturing and sending electronic documents in exactly the intended format.
- Unlike Microsoft Word and Excel, key challenge with PDF format is extracting and editing information.
- OCR (Optical Character Recognition) technique is used to identify words in a picture/scanned document and convert it into a machine readable text, that can be processed further with the help of computer.
- Although the technology is matured and uses advanced techniques it quite often produces erroneous output.

Breaking down the problems

Problem Definition: Identification of specified OCR errors in the document and correction of the same

01 Identification of the OCR errors in a text document

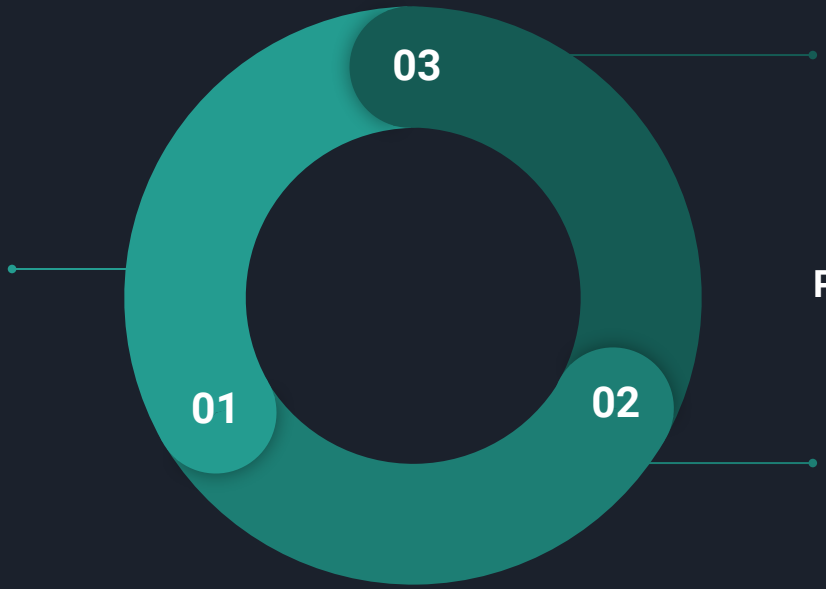
02 De-Noising of OCR-Spelling correction/replacing wrong character with the right character

03 Generate respective DOCX file for every PDF resolved

Solution Design

OCR PDF Files

- Using Poppler run OCR on the provided PDF files
- Generate XML file corresponding to every Optically Read PDF file.
- The XML stores the textual, and syntactical data of the PDF.



Create DOCX

- OCR results are not assured to provide data in right-left and top-down order.
- Above problem resolved
- Use python-docx library to create DOCX file as closely resembling the original PDF

Parse XML

- Parse the XML file to resolve OCR errors i.e, De-Noise OCR
- Commonly encountered OCR mistakes and spell checking implemented
- 2D List with multiple nested dictionaries data structure implemented to better store, understand and reuse XML content.

Additionally:
Multiprocessing with batch processing support is implemented execute the solution for multiple documents parallely

Execution Steps

Step 1 Preprocessing

- Place all PDF documents in PDFs folder.
- Install all pre requisite tools needed
- Execute main.py

Step 2

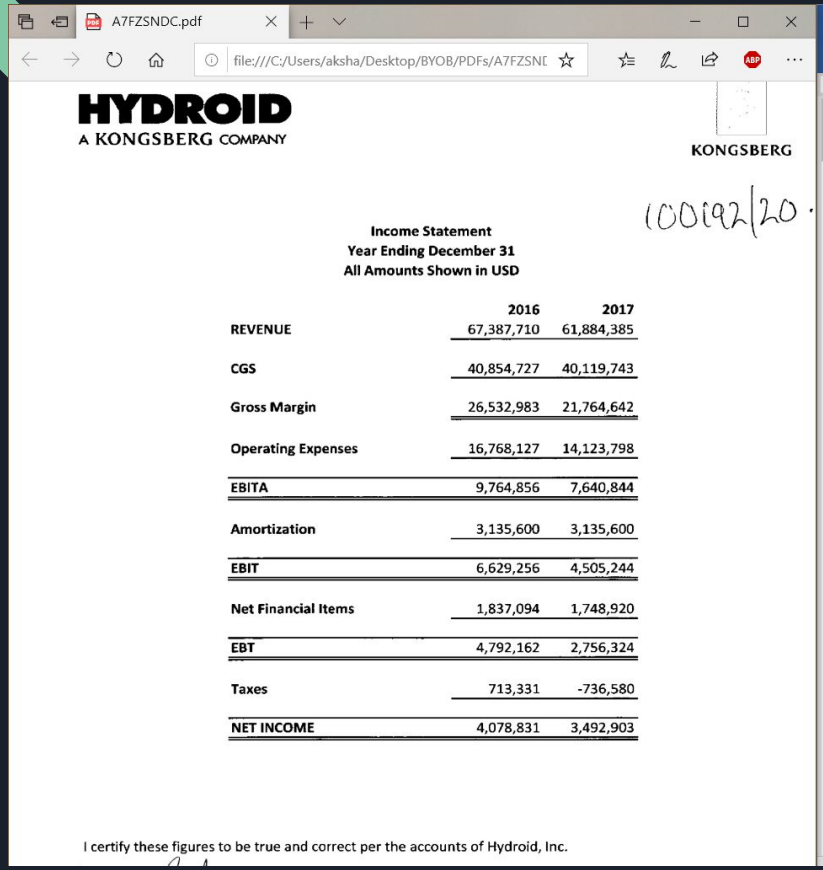
- Respective DOCX files are generated in DOCX folder with subfolder of filename of PDF

Step 3

- View and analyse the Results
- Achieved good accuracy in OCR reading and De-Noising
- Only minor issues in docx conversion exist

Few Obtained Results

Side by Side Comparisons
Right: PDF | Left: Converted Docx



A7FZSNDC.pdf

file:///C:/Users/aksha/Desktop/BYOB/PDFs/A7FZSNI

HYDROID
A KONGSBERG COMPANY

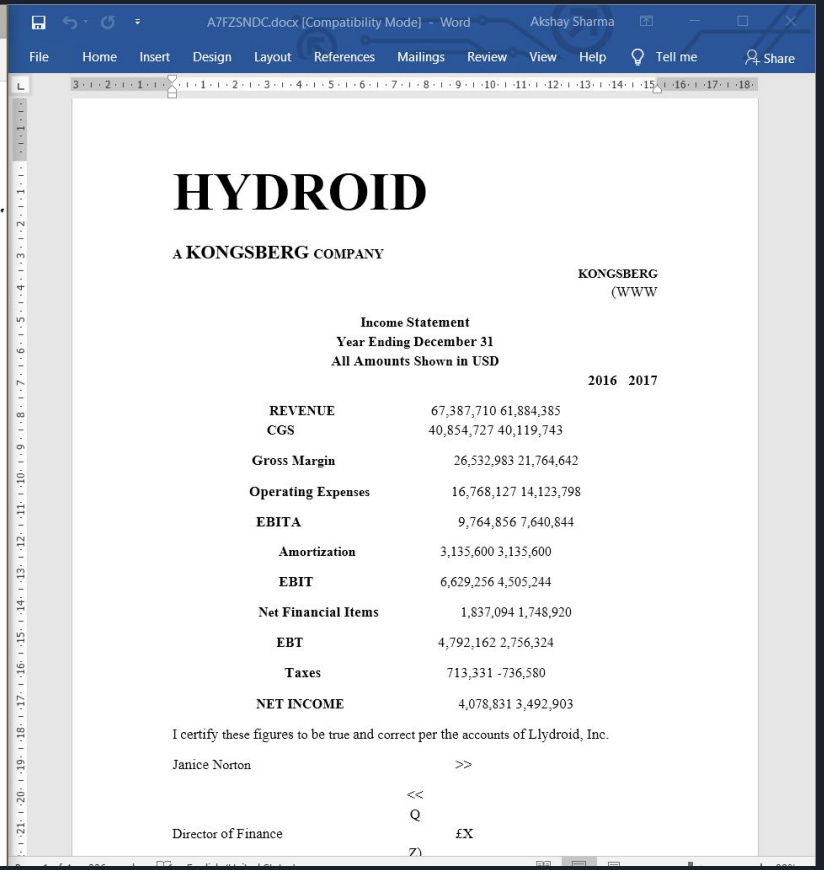
KONGSBERG

100192|20

Income Statement
Year Ending December 31
All Amounts Shown in USD

	2016	2017
REVENUE	67,387,710	61,884,385
CGS	40,854,727	40,119,743
Gross Margin	26,532,983	21,764,642
Operating Expenses	16,768,127	14,123,798
EBITA	9,764,856	7,640,844
Amortization	3,135,600	3,135,600
EBIT	6,629,256	4,505,244
Net Financial Items	1,837,094	1,748,920
EBT	4,792,162	2,756,324
Taxes	713,331	-736,580
NET INCOME	4,078,831	3,492,903

I certify these figures to be true and correct per the accounts of Hydroid, Inc.



A7FZSNDC.docx [Compatibility Mode] - Word

Akshay Sharma

File Home Insert Design Layout References Mailings Review View Help Tell me Share

HYDROID

A KONGSBERG COMPANY

KONGSBERG
(WWW

Income Statement
Year Ending December 31
All Amounts Shown in USD

	2016	2017
REVENUE	67,387,710	61,884,385
CGS	40,854,727	40,119,743
Gross Margin	26,532,983	21,764,642
Operating Expenses	16,768,127	14,123,798
EBITA	9,764,856	7,640,844
Amortization	3,135,600	3,135,600
EBIT	6,629,256	4,505,244
Net Financial Items	1,837,094	1,748,920
EBT	4,792,162	2,756,324
Taxes	713,331	-736,580
NET INCOME	4,078,831	3,492,903

I certify these figures to be true and correct per the accounts of Llydroid, Inc.

Janice Norton >>

<<
Q

Director of Finance £X

Z)

A7GZKX9L.pdf

file:///C:/Users/aksha/Desktop/BYOB/PDFs/A7GZKX9

11 ST. CROSS STREET LIMITED

BALANCE SHEET

AS AT 31 JANUARY 2018

	2018		2017	
	£	£	£	£
Fixed assets		1,428,390		1,428,390
Creditors: amounts falling due within one year	(420,951)		(386,031)	
Net current liabilities		(420,951)		(386,031)
Total assets less current liabilities		1,007,439		1,042,359
Creditors: amounts falling due after more than one year		(998,231)		(1,035,554)
Net assets		9,208		6,805
Capital and reserves		9,208		6,805

Notes to the financial statements

1 Average employees
The average number of persons (including directors) employed by the company during the year was 0 (2017 - 0).

2 Group Accounts
The financial statements present information about the company as an individual undertaking and not about its group. The company and its subsidiary undertaking comprise a small-sized group. The company has therefore taken advantage of the exemptions provided by section 399 of the Companies Act 2006 not to prepare group accounts.

3 Bank Loan
The mortgage of £1,037,120 (2017: £1,072,097) is secured by a charge over property held within the subsidiary, by personal guarantees given by the directors and by a floating charge over the assets of the company.

11 St. Cross Street Limited is a private company limited by shares incorporated in England and Wales. The registered office is Kenton House, 666 Kenton Road, Harrow, Middlesex, HA3 9QN.

The directors of the company have elected not to include a copy of the profit and loss account within the

A7GZKX9L.docx [Compatibility Mode] - Word — Akshay Sharma

File Home Insert Design Layout References Mailings Review View Help Tell me Share

11 ST. CROSS STREET LIMITED

BALANCE SHEET

AS AT 31 JANUARY 2018

	2018		2017	
	£	£	£	£
Fixed assets		1,428,390		1,428,390
Creditors: amounts falling due within one year	(420,951)		(386,031)	
Net current liabilities		(420,951)		(386,031)
Total assets less current liabilities		1,007,439		1,042,359
Creditors:, amounts falling due after more than one year		(998,231)		(1,035,554)
Net assets		9,208		6,805
Capital and reserves		9,208		6,805

Notes to the financial statements

1 Average employees
The average number of persons (including directors) employed by the company during the year was 0
(2017-0).
2 Group Accounts
The financial statements present information about the company as an individual undertaking and not
about its group. The company and its subsidiary undertaking comprise a small-sized group. The company
has therefore taken advantage of the exemptions provided by section 399 of the Companies Act 2006 not
to prepare group accounts.
3 Bank Loan

Page 2 of 5 466 words English (United States) 90%

A7H4LIP7.pdf

file:///C:/Users/aksha/Desktop/BYOB/PDFs/A7H4LIP

PARITUM PROJECTS LLP

United Kingdom, Registered Company No. 0C321042

REPORT OF THE MEMBERS FOR THE FINANCIAL PERIOD FROM 1.08.2017. TO 31.07.2018.

The Members herewith present their report with the Financial Statements of the Company for the period ended 31 July 2018.

INCORPORATION & REGISTERED OFFICE

The Company incorporated on 18 July 2006 under Registered Company Number 0C321042. The registered office address is at Unit 5, Olympia Industrial Estate, Coburg Road, London, England, N22 6TZ.

PRINCIPAL ACTIVITY & REVIEW OF BUSINESS

During the current financial period the company was active as trade agent for construction materials and received commission.

NOTE

The company primarily conducts business in foreign currencies, such as Euros. These transactions have been converted into Pounds Sterling at averaged Exchange rates throughout the current financial year period.

RESPONSIBILITY

The members are required by Company Law to prepare accounts for each financial year which give a true and fair view of the state of affairs of the Company as at the end of the financial year under review and of the result of that period and deliver the same to the registrar.

The members of the partnership are responsible for maintaining adequate accounting records and for the safeguarding of the books and assets of the Company.

KANG DONG-HEE for
FORMOND INC.
Designated member

Approved by the members on 12 September 2018.

A7H4LIP7.docx [Compatibility Mode] - Word

Akshay Sharma

File Home Insert Design Layout References Mailings Review View Help Tell me Share

PARITUM PROJECTS LLP

United Kingdom, Registered Company No. 0C321042

REPORT OF THE MEMBERS FORTHE FINANCIAL PERIOD

FROM 1.08.2017. TO 31.07.2018.

The Members herewith present their report with the Financial Statements of the Company for the

period ended 31 July 2018.

INCORPORATION && REGISTERED OFFICE

The Company incorporated on 18 July 2006 under Registered Company Number 0C321042. The registered office address is at Unit 5, Olympia Industrial Estate, Coburg Road, London, England, N22 6TZ.

PRINCIPAL ACTIVITY && REVIEWOF BUSINESS

During the current financial period the company was active as trade agent for construction materials and received commission.

NOTE

The company primarily conducts business in foreign currencies, such as Euros. These transactions have been converted into Pounds Sterling at averaged Exchange rates throughout the current financial year period.

RESPONSIBILITY

The members are required by Company Law to prepare accounts for each financial year which give a true and fair view of the state of affairs of the Company as at the end of the financial year under review and of the result of that period and deliver the same to the registrar.

The members of the partnership are responsible for maintaining adequate accounting records and for the safeguarding of the books and assets of the Company.

KANG DONG-HEE for

FORMOND INC.
Designated member
Approved by the members on 12 September 2018.

COMPANIES HOUSE

Any Motes form part of these Financial Statements

Page 1 of 5 551 words English (United States) 90%

Thank you!