

# Qarik Group Corporate Project

## Group Bear

K. Bari   K. Bombardier   A. Castano   B. Liu

Erdos Institute Presentations, Fall 2021

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis
  - Visualization
  - Clustering
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work

# Examples of Loan Document

Public Disclosure Authorized

Public Disclosure Authorized

CONFORMED COPY

LOAN NUMBER 3247 POL

(Structural Adjustment Loan)

between

REPUBLIC OF POLAND

and

INTERNATIONAL BANK FOR RECONSTRUCTION  
AND DEVELOPMENT

Dated August 31, 1990

LOAN NUMBER 3247 POL

LOAN AGREEMENT

AGREEMENT, dated August 31, 1990, between REPUBLIC OF POLAND  
(the Borrower) and INTERNATIONAL BANK FOR RECONSTRUCTION AND  
DEVELOPMENT (the Bank).

WHEREAS (a) the Bank has received a letter dated July 2,  
1990, from the Borrower describing a program of actions, objectives  
and policies designed to achieve structural adjustment of the  
Borrower's economy (hereinafter called the Program), declaring the  
Borrower's commitment to the execution of the Program, and request-  
ing assistance from the Bank in the financing of urgently needed  
imports and services required during such execution; and

(b) on the basis, inter alia, of the foregoing, the Bank has  
decided in support of the Program to provide such assistance to the  
Borrower by making the Loan in two tranches as hereinafter provided;

NOW THEREFORE the parties hereto hereby agree as follows:

## ARTICLE I

### General Conditions/Definitions

Section 1.01. The "General Conditions Applicable to Loan and  
Guarantee Agreements" of the Bank, dated January 1, 1985, with the  
modifications thereto set forth below (the General Conditions)  
constitute an integral part of this Agreement:

(a) Section 2.01, paragraph 11, shall be modified to read:

"'Project' means the imports and other activities that  
may be financed out of the proceeds of the Loan pursuant to  
the provisions of Schedule 1 to the Loan Agreement,"

(b) Section 2.07 (c) shall be modified to read:

"Not later than six months after the closing date of  
such later date as may be agreed for this purpose between the  
Borrower and the Bank, the Borrower shall prepare and furnish  
to the Bank a report, of such scope and in such detail as the  
Bank shall reasonably request, on the execution of the Program  
referred to in the Preamble to the Loan Agreement; the  
performance by the Borrower and the Bank of their respective  
obligations under the Loan Agreement and the accomplishment of  
the purposes of the Loan," and

(c) The last sentence of Section 3.02 is deleted.

Section 1.02. Unless the context otherwise requires, the  
several terms defined in the General Conditions and in the Preamble  
to this Agreement have the respective meanings therein set forth and  
the term "BITC" means the Standard International Trade  
Classification, Revision 3 (SITC, Rev. 3), published by the United  
Nations in Statistical Papers, Series M, No. 343 (1981).

## ARTICLE II

### The Loan

Section 2.01. The Bank agreed to lend to the Borrower, on the  
terms and conditions set forth or referred to in the Loan Agreement,  
various currencies that shall have an aggregate value equivalent to  
the amount of three hundred million dollars (\$300,000,000), being  
the sum of withdrawals of the proceeds of the Loan, with each  
withdrawal valued by the Bank as of the date of such withdrawal.

Section 2.02. The amount of the Loan may be withdrawn from  
the Loan Account in accordance with the provisions of Schedule 1 to  
this Agreement.

Section 2.03. The Closing Date shall be December 31, 1991 or  
such later date as the Bank shall establish. The Bank shall promptly  
notify the Borrower of such later date.

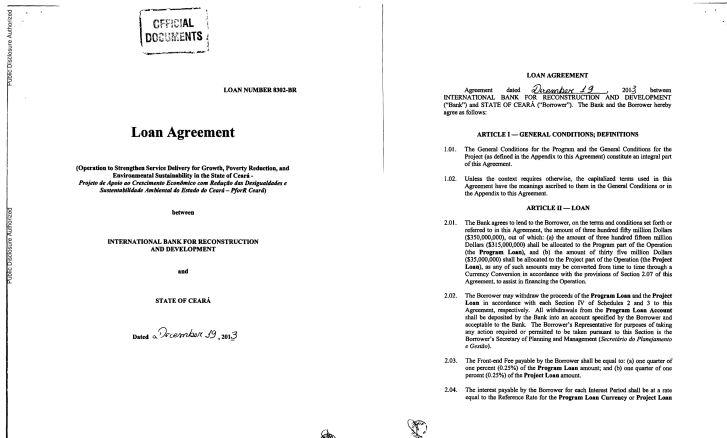
Section 2.04. The Borrower shall pay to the Bank a commitment  
charge at the rate of three-fourths of one percent (3/4 of 1%) per  
annum on the principal amount of the Loan not withdrawn from time to  
time.

Section 2.05. (a) The Borrower shall pay interest on the  
principal amount of the Loan withdrawn and outstanding from time to  
time, at a rate for each interest period equal to the cost of  
Qualified Borrowings determined in respect of the preceding  
semester, plus one-half of one percent (1/2 of 1%). On each of the  
dates specified in Section 2.06 of this Agreement, the Borrower  
shall pay interest accrued on the principal amount outstanding  
during the preceding interest period, calculated at the rate  
applicable during such interest period.

(b) As soon as practicable after the end of each semester,  
the Bank shall notify the Borrower of the Cost of Qualified  
Borrowings determined in respect of such semester.

# Examples of Loan Document

## Scanned PDFs



# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis
  - Visualization
  - Clustering
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work

# Extraction Tools

Python Packages considered:

- PDFMiner.six – Fast, cannot read scanned PDFs, section structure not preserved
- PyMuPDF – Fast, cannot read scanned PDFs
- pyTesseract – Good for scanned PDFs, computationally expensive, OCR inaccurate

# Results of Extraction

Total Number of Documents: 3205 ranging from 1990 to 2019

- PyMuPDF: 2805 files scanned accurately
- pyTesseract: 400 remaining files

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis
  - Visualization
  - Clustering
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work



# Features extracted from PDFs

- Date (100%)
- Country (100%)
- Loan Amount (and currency) (92%)
- Project Name
- Project Description (91%)

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis
  - Visualization
  - Clustering
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work

# Features extracted from External Sources

- GDP
- Historic Exchange Rates

# Examples of Loan Document

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 **Data Analysis**
  - **Visualization**
  - Clustering
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis**
  - Visualization
  - Clustering**
  - Analysis of Wealth and Loan information
- 5 Conclusions and Future Work

# Outline

- 1 Text Extraction
  - PyMuPDF and pyTesseract
- 2 Data Extraction
  - Data from PDFs
  - Data from other sources
- 3 Key Performance Indicators
- 4 Data Analysis**
  - Visualization
  - Clustering
  - **Analysis of Wealth and Loan information**
- 5 Conclusions and Future Work

# Make Titles Informative. Use Uppercase Letters.

Subtitles are optional.

- Use `itemize` a lot.
- Use very short sentences or short phrases.



# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

# Make Titles Informative.

You can create overlays. . .

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.



Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

Text Extraction  
Data Extraction  
Key Performance Indicators  
Data Analysis  
Conclusions and Future Work  
Summary

# Make Titles Informative.

# Summary

- The **first main message** of your talk in one or two lines.
- The **second main message** of your talk in one or two lines.
- Perhaps a **third message**, but not more than that.
- Outlook
  - Something you haven't solved.
  - Something else you haven't solved.

# For Further Reading I



A. Author.

*Handbook of Everything.*

Some Press, 1990.



S. Someone.

On this and that.

*Journal of This and That*, 2(1):50–100, 2000.