

# Qarik Group Corporate Project

## Group Bear

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

Erdos Institute Presentations, Fall 2021

# Outline

- 1 Text Extraction
  - Examples
  - 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
  - Examples
    - 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

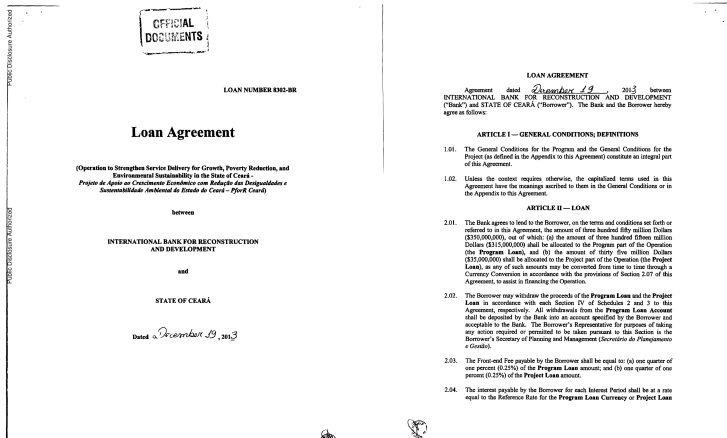
Public Disclosure Authorized

(10) 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.

NOW THEREFORE the parties hereto hereby agree as follows:

# Examples of Loan Document

## Scanned PDFs



# Outline

- 1 Text Extraction
  - Examples
  - 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
  - Examples
  - 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
  - Examples
  - 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
  - Examples
  - 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
  - Examples
  - 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
  - Examples
  - 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.