Qarik Group Corporate Project Group Bear

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

Erdos Institute Presentations, Fall 2021



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



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



Examples

PyMuPDF and pyTesseract

Examples of Loan Document

INTERNATIONAL BANK FOR RECONSTRUCTION

LOW STMEEN 3247 FOL LONG MISSESSORT

ASSEMBLY, dated August 31, 1990, between SEPUBLIC OF POLAND (the Berrower) and INTERNATIONAL BASK FOR RECONSTRUCTION AND DEVELOPMENT (the Bask).

SWEEDER (A) the Bark has received a letter dated July 2.
100, from the Norreser describing a program of actions, dejectives
and policies designed to achieve necessaries lettered. Editories of the
Borrower's accommonly hereinforther called the Frogram, declaring the
Borrower's commitment to the execution of the Frogram, and requesting assistance from the Bark in the financing of unpertyl smedded

(8) on the basis, inter alia, of the foregoing, the Bank has decided in support of the Frogram to provide outh asolatence to the Borrower by making the Loan in two tranches as hereinafter provided; NOW THEREFORE the parties hereto hereby agree as follows:

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 thereof set feeth below (the General Conditions) constitute an internal part of this Agreement:

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

(B) Section 9.07 (c) shall be modified to read "(c) Not later than six mosths after the Closing Date or

"(c) but later than six mouths after the cleaning mate or such later date as may be spread for this purpose between the Borrower and the Bank, the Surrower shall prepare and furnish to the Bank a request, of such moves and in order detail as the prepared to be the resulting the such as the such as the referred to be the Personality to the Lank Agreement, the performance by the Borrower and the Bank of their respective obligations under the Lank Agreement and the accomplishment of the pumpers of the Lount, Yau

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

Rection 1.02. Unless the context otherwise requires, the several terms diefined in the General Conditions and in the Treamble the term fully mass the Charded International Teach Classification, Revision 3 (SITC, New. 3), poblished by the United National In Extincted Impers, Derice N. 80, 341 (1989)

ANTICLE II The Loan

section 3.01. The Bank agreem to lead to the Morrower, on the torms and scodilions set forth or referred to in the Lean Agreement, various of the second section of the second section of the amount of three humaned million solians (\$100,000,001), being the sec of withdrawals of the proceeds of the Lean with each withdrawal valued by the Mank as of the date of such withdrawal.

Section 2.06. The Borrower shall pay to the Bank a commitment charge at the rate of three-fourths of one percent (574 of 28) per answe on the principal amount of the loan not withdraws from time to

shall pay interest accreed on the principal amount constanding during the precedular laterest Feriod, saloulated at the rote applicable during such Interest Pariod. (8) As some appreciabable after the end of each Semester, the Bank shall notify the Secrewar of the Cost of Gmalified Secreeings decembed in respect of such Semester.

Examples

PyMuPDF and pyTesseract

Examples of Loan Document

Scanned PDFs



LOAN AGREEMENT

Agreement dated Accounting 19, 2013 between INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT ("Bark") and STATE OF CEARA ("Berrower"). The Bank and the Berrower beredy agree as follows:

ARTICLE I - GENERAL CONDITIONS; DEFINITIONS

- 1.01. The General Conditions for the Program and the General Conditions for the Project (as defined in the Appendix to this Agreement) constitute an integral part of this Agreement.
- 1.02. Unless the context requires otherwise, the capitalized terms used in this Agreement have the meanings secribed to them in the General Conditions or in the Agrendix to this Agreement.

ARTICLE II - LOAN

- 2.10. The Back agrees to find to the Roomers, on the terms and conditions set father enferred to in this Agreement, the remote of their handred light william Dollars (335500,00200), out of which (a) the amount of three handred fithern million Dollars (3315,000,000), and of which (a) the amount of three handred fither million Dollars (3315,000,000) and the allocated to the Propers are for the Operation (the Propers Loue), and (b) the amount of three five five million Dollars (335,000,000) all he allocated to the Propers are of the Operation (a) Propers (a) Propers (b) Propers (c) Prope
- 2.02. The Borrower may windraw the proceeds of the Program Loss and the Project Loss in accordance with each Section [V of Schoduler 2 and 3 to this Agreement, respectively. All windraword from the Program Loss Account shall be despoted by the Barrower and the despoted by the Barrower and the section of the Comparison of the Barrower's Section (as expected or permitted to be taken paramet to this Section in the Berrower's Sectionary of Planning and Management (Serverisch of Programsome).
- 2.03. The Front-end Fee payable by the Borrower shall be equal to: (a) one quarter of one percent (0.25%) of the Program Lass amount; and (b) one quarter of one percent (0.25%) of the Project Lass amount.
- 2.04. The interest payable by the Borrower for each Interest Period shall be at a rate equal to the Reference Rate for the Program Loan Currency or Project Loan





- Text Extraction
 - Examples
 - PyMuPDF and pyTesseract
- 2 Data Extraction
 - Data from PDFs
 - Data from other sources
- Key Performance Indicators
- Data Analysis
 - Visualization
 - Clustering
 - Analysis of Wealth and Loan information
- 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

Examples
PyMuPDF and pyTesseract

Results of Extraction

Total Number of Documents: 3205 ranging from 1990 to 2019

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

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



Features extracted from PDFs

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

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



Features extracted from External Sources

- GDP
- Historic Exchange Rates

Examples of Loan Document

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



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



- 1 Text Extraction
 - Examples
 - PyMuPDF and pyTesseract
- Data Extraction
 - Data from PDFs
 - Data from other sources
- Key Performance Indicators
- 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.

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



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



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



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



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



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



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.