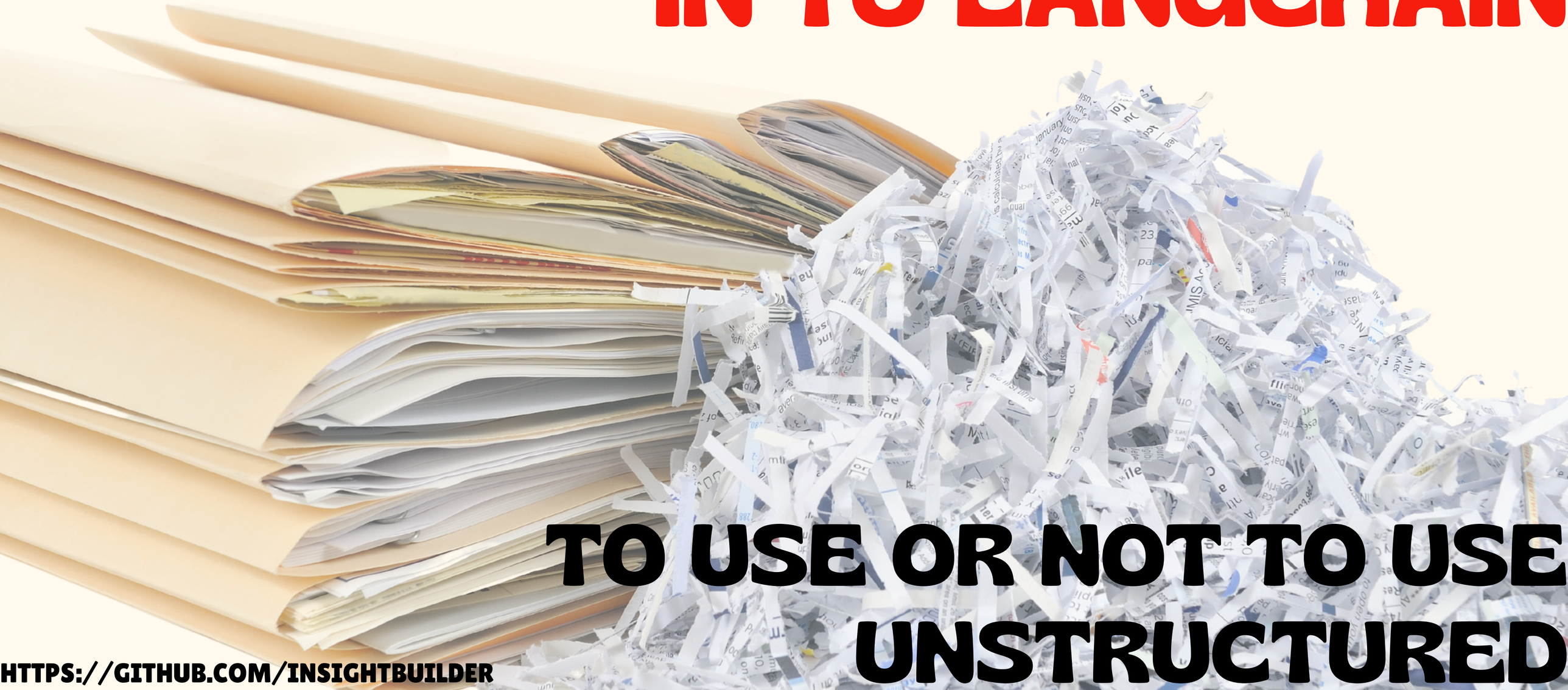
LOADING PDFS (DATA) IN TO LANGCHAIN



TO USE OR NOT TO USE UNSTRUCTURED

HTTPS://GITHUB.COM/INSIGHTBUILDER

MULTIPLE CHOICES : SAME CHALLENGE

- PYPDF
- UNSTRUCTUREDPDFLOADER
- PDFMINER
- PYMUPDF

WHY SO MANY CHOICES?

PDFS DOCUMENTS ARE USED FOR DIFFERENT PURPOSES. DEPENDING ON YOUR APPLICATION YOU MAY CHOOSE THE ONE BEST SUITS YOU.

IT ALL BOILS DOWN TO : HOW THE PDFS ARE SHREDDED

PDFS DOCUMENTS ARE USED FOR DIFFERENT PURPOSES. DEPENDING ON YOUR APPLICATION THE PDF LAYOUT WILL BE DIFFERENT. THE CHOICE OF WORK WITH IMAGES/ FIGURES NEEDS TO THOUGHT OUT

BEFORE ENTERING THE RABBIT HOLE

UNSTRUCTURED[LOCAL-INFERENCE] :

USE IT ONLY YOU ARE DOING

1) ADVANCED PDF PARSING

2) PDF FROM MULTIPLE SOURCES

3) HAVE LOT OF TIME

4) READY TO EXPLORE

PYPDF :

IF IN HURRY, STILL CHOOSE THE RED PILL...

JUST KIDDING. USE PYPDF, ALONG WITH RECURSIVE TEXT SPLITTER

HTTPS://GITHUB.COM/INSIGHTBUILDER

GOING DOWN THE RABBIT HOLE

UNSTRUCTURED[LOCAL-INFERENCE] : USES MACHINE LEARNING MODELS UNDER THE HOOD TO WORK ON THE PDF, IMAGE FILES. HAVING THE MODELS LOCALLY WILL HELP TO PROCESS PDF BETTER

- PARTITION_PDF
- PARTITION_EMAIL
- PARTITION_HTML
- PARTITION

PYPDF : OPEN SOURCE PURE-PYTHON PDF LIBRARY CAPABLE OF SPLITTING, MERGING, CROPPING, AND TRANSFORMING THE PAGES OF PDF FILES.

- PDFREADER CLASS
- PDFWRITER CLASS
- PDFMERGER CLASS
- PAGEOBJECT CLASS

UNSTRUCTURED : AN INTERFACE

- PARTITION CURRENTLY SUPPORTS .DOCX, .DOC, .PPTX, .PPT, .EML, .MSG, .EPUB, .HTML, .PDF, .PNG, .JPG, AND .TXT FILES.

ELEMENT SUPPORT

- TEXT
- FIGURECAPTION
- NARRATIVETEXT
- LISTITEM
- TITLE
- ADDRESS
- PAGEBREAK
- CHECKBOX
- IMAGE

ML INTEGRATIONS

- ARGILLA
- HUGGING FACE
- DATASAUR
- LABELBOX
- LABEL STUDIO
- LANGCHAIN
- PANDAS
- PRODIGY

BRICKS

- PARTITION : MAJOR WORK
- CLEANING : TEXT CLEAN
- STAGING : CONNECTING

STAGING:

- CONVERT_TO_DICT
- DICT_TO_ELEMENTS
- CONVERT_TO_CSV
- CONVERT_TO_DATAFRAME
- STAGE_FOR_TRANSFORMERS

NEXT SEE WHAT LANGCHAIN ABSTRACTION DOES TO THE PDF

- WILL SPLIT LIKE UNSTRUCTED

- WILL ATTACH METADATA AND MAKE IT A DOCUMENT

LETS HEAD TO COLAB

NOW WE ARE TALKING!!! PRACTICE

THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER