THINK BEFORE YOU SHRED IT

FEEDING PDFS (DATA) TO LANGUAGE MODELS

HTTPS://GITHUB.COM/INSIGHTBUILDER

ITS ABOUT CHALLENGES

ALWAYS REMEMBER LIBRARIES EXIST TO SOLVE SMALL CHALLENGES

TAKE A CHALLENGE

LEARN ABOUT IT AND BREAK THE CHALLENGE INTO SMALLER CHALLENGES << THIS STEP IS KEY

THEN RESEARCH ABOUT THE TOOLS THAT YOU CAN USE TO SOLVE THEM.

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

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

THE ML MODELS ARE PULLED INTO YOUR SYSTEM THROUGH PIP INSTALLS ARE APT INSTALLS.

- LIBMAGIC-DEV (FILETYPE DETECTION)
- POPPLER-UTILS (IMAGES AND PDFS)
- TESSERACT-OCR (IMAGES AND PDFS)
- LIBREOFFICE (MS OFFICE DOCS)

- PARTITION_PDF
- PARTITION_EMAIL
- PARTITION_HTML
- PARTITION

PIP INSTALL "DETECTRON2@GIT+HTTPS://GITHUB.COM/FACEBOOKRESEARCH/DETECTRON2.GIT@E2CE8DC#EGG-DETECTRON2

HTTPS://GITHUB.COM/INSIGHTBUILDER

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

UNDERSTAND HOW UNSTRUCTURED WORKS WITH THE PDF.

NEXT SEE WHAT LANGCHAIN ABSTRACTION DOES TO THE PDF

- WILL SPLIT LIKE UNSTRUCTED
- WILL ATTACH METADATA AND MAKE IT A DOCUMENT

DECIDE HOW TO PROCEED FROM THERE.

DO I NEED UNSTRUCTURED?

SOMETIMES IT IS BETTER TO WORK WITH BIGGER CHUNK AND THEN SPLIT IT WITH ANOTHER SPLITTER

LETS HEAD TO COLAB

NOW WE ARE TALKING!!! PRACTICE

THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER