**查看这两个数据集的技术文档**

**熟悉两个数据集里面所有字段的含义**

**下载两个数据集的一部分数据尝试加载**

**这两个数据集有什么差异**

**可视化位置信息在pdf上**

* **pixparse/pdfa-eng-wds数据集：**

Pdfs-eng-tain-xxxx.tar文件解压之后内容：包含pdf详细信息的json文件与匹配的pdf文件

**1.\_\_url\_\_属性: 这一个pdf文件在哪个压缩包的路径**

**2.json属性:**

Pages内容： 包括不同页面的信息content的列表

每个content都是一个dict:

      1.words属性： text包含了不带空格的单词列表；bbox包含每个单词与页面大小相关的左、顶、宽、高格式的边界框坐标；Line\_pos是一个元组列表，它表示单词所属行的索引，然后是该行中字符的起始位置;score值为1.0的占位符

      2.lines属性：  text包含了启发式方法组合在一起的线（序列的部分，由空格分隔的字符串，按阅读顺序组织重新塑造）；bbox包含与页面大小相关的左、顶、宽、高格式的边界框坐标;score值为1.0的占位符；word\_slice没有指出有什么意义

      3. images\_bbox为list类型：嵌入到页面中的图像的边界框

      4.Images\_bbox\_no\_text\_overlap为list类型:     与pdf中发现的文本没有重叠的边界框的简化列表

**3.pdf属性：pdf文件内容**

* **pixparse/idl-wds数据集：**

**1.\_\_url\_\_属性:** 这一个pdf文件在哪个压缩包的路径

**2.ocr属性:**pdf原始的、较旧的OCR注释

**3.json属性:**  pdf使用idl\_data项目的大量文本OCR注释（使用Amazon的OCR引擎text来获取OCR注释）

pages属性：文档中每个页面content的列表

每个content内容：

——text属性：包括按照阅读顺序组织每一行具体文本的列表

——bbox属性：每一行的包含与页面大小相关的左、顶、宽、高格式的边界框坐标

——poly属性: 每一行的左上，右上，右下，左下的坐标

——score属性： 每一行文本获取的置信度分数

4.**pdf属性：**pdf文件内容

5.**tif属性 ：**pdf文件的图片形式