Atitit  信息管理 艾提拉著作 CAPT3 数据归一化 文本化 与结构化

目录

[1. 文本化内容 2](#_Toc10806)

[1.1. 文件列表 2](#_Toc28510)

[1.2. 文件内容 2](#_Toc15068)

[2. 2. 常见的类型 2 2](#_Toc25077)

[2.1. 2.1.1. Type application 2 2](#_Toc158)

[2.2. 2.1.4. Type text 3 2](#_Toc5818)

[2.3. 2.1.3. Type image 3 2](#_Toc13390)

[2.4. 2.1.2. Type audio 2 2](#_Toc22857)

[2.5. 2.1.5. Type video 4 2](#_Toc22758)

[3. 1. Step1.从压缩文件抽取 2 3](#_Toc5975)

[3.1. 1.1. Zip文档 2 1.2. Rar文档 2 3](#_Toc18780)

[3.2. 1.3. 复合压缩文档 2 3](#_Toc15281)

[4. 2. Step2.从文档抽取纯文本 office pdf eml html文档 2 3](#_Toc6711)

[4.1. 2.1. office （word ppt excel 2 2.2. Pdf 2 3](#_Toc12808)

[4.2. 2.3. Eml 2 2.4. Html mhtml 2 3](#_Toc5274)

[4.3. 2.5. 图片 ocr（jpg png 2 3](#_Toc25153)

[5. 3. 目录提取+ Keyword提取 2 3](#_Toc32485)

[6. 数据结构化 3](#_Toc10018)

[6.1. 目录tag化 3](#_Toc18202)

[6.2. 文件tag 3](#_Toc9112)

[6.3. 压缩包tag 3](#_Toc1443)

[6.4. Excel结构化 3](#_Toc13025)

[6.5. 扩展属性 比如mp3 扩展年代 容易程度 男女歌手 情感等 4](#_Toc17210)

# 文本化内容

## 文件列表

## 文件内容

# 2. 常见的类型 2

## 2.1.1. Type application 2

## 2.1.4. Type text 3

## 2.1.3. Type image 3

## 2.1.2. Type audio 2

## 2.1.5. Type video 4

# 1. Step1.从压缩文件抽取 2

## 1.1. Zip文档 2 1.2. Rar文档 2

## 1.3. 复合压缩文档 2

# 2. Step2.从文档抽取纯文本 office pdf eml html文档 2

## 2.1. office （word ppt excel 2 2.2. Pdf 2

## 2.3. Eml 2 2.4. Html mhtml 2

## 2.5. 图片 ocr（jpg png 2

# 3. 目录提取+ Keyword提取 2

# 数据结构化

## 目录tag化

## 文件tag

## 压缩包tag

## Excel结构化

## 扩展属性 比如mp3 扩展年代 容易程度 男女歌手 情感等

Atitit  信息管理 艾提拉著作 CAPT4 数据tag标注与结构化