外部数据导入WordPress方法的初探

# WordPress

WordPress使用GPL 授权发布许可，可以一直免费使用。

最近安装了WordPress 4.9.8

安装在/opt/wordpress目录下。Wordpress 4.9.8（bitanami安装版）

# 将Excel中的数据导入WordPress

## Xlrd库

在python中可以使用xlrd从excel中读出数据。

pip install xlrd

我写了一个例子程序，从excel中读出数据，然后插入WordPress的数据库。

excel\_to\_wordpress.py

数据源在以下目录：

/home/jankin/examples/

excel\_to\_wordpress.py使用以下库：

|  |  |
| --- | --- |
| **库** | **用途** |
| xlrd | 读取excel数据 |
| MySQLdb | 将数据插入Mysql数据库。  WordPress使用Mysql数据库来存储数据。 |
| wordpress\_xmlrpc | 在WordPress中创建新的提交。 |

excel\_to\_wordpress.py只完成了部分工作，excel中的数据被读出，根据数据在WordPress中创建了Post，然后为Post补充了一些属性（将一些属性插入到wp\_postmeta），但是Post还是在草稿箱里，没有发布。

为了方便发布这些文章，我把创建的文章的ID保存以下文件中：

/opt/tmp/postId.txt

## 批量发布Post

如果想发布Post，需要用root账号。

cd /opt/wordpress-4.9.8-0/apps/wordpress

cat /opt/tmp/postId.txt | xargs –t –I {} bin/wp post update {} --post\_status=publish --allow-root

保存在文件里的post id也可以方便的来删除不想要的Post。

比如：

/opt/tmp/post\_id\_pdf.txt中的这些Post，是用pdfminer从PDF文档中导出的数据创建的。但是这个数据不全，有顺序错误，也缺少图片，所以这些Post被删除了。

cat /opt/tmp/post\_id\_pdf.txt | xargs –t –I {} bin/wp post delete {} --allow-root

# 将PDF中数据导入到WordPress

## 用pdfminer读取PDF文档

实践证明，pdfminer库是可以将PDF文档中的数据，但是读出的数据的顺序不正确，特别是读取PDF中的表格数据时，问题尤其严重。

另外用pdfminer也没法把PDF中数据中的图片导出。

## 用poppler-utils将PDF文档中数据导出

poppler-utils效果更好。

在Ubuntu上用以下命令安装poppler-utils：

sudo apt-get install poppler-utils

为了方便处理，可以将PDF中的内容输出为xml文件和一组图片。

pdftohtml –xml 1.pdf 1.xml

执行完上述命令后，除了生成1.xml文件，还在同一目录中生成了PDF中图片。

## xml.sax

在python中使用xml.sax库可以轻松的将xml变成自己的内存结构。

我写了一个例子程序，从xml中读出数据，然后插入WordPress的数据库。

xml\_to\_wordpress.py

这个程序只是负责根据xml中的数据创建Post。图片需要单独上传到WordPress

使用以下命令将目录中的所有图片导入到WordPress中。

cd /opt/wordpress-4.9.8-0/apps/wordpress

bin/wp media import /home/jankin/example/xml/1/\*.jpg --allow-root

xml\_to\_wordpress.py创建的Post的id保存到以下文件中：

/opt/tmp/post\_id\_request.txt