字符生成模型

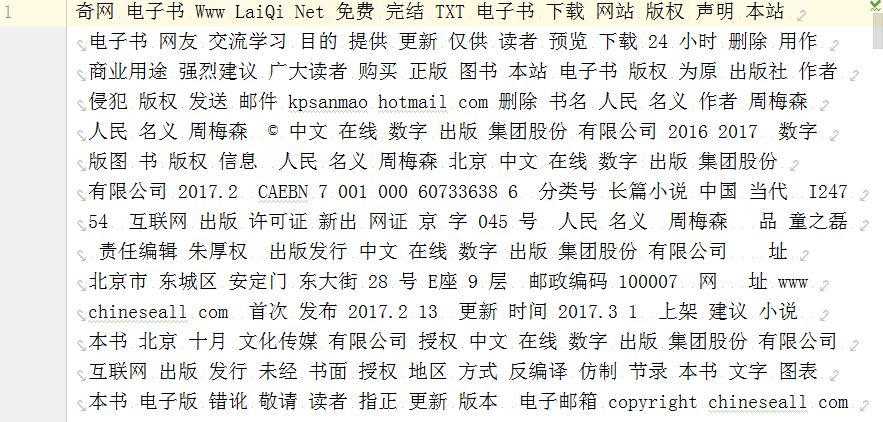
1. 字符生成模型介绍

该字符生成模型用于生成SynthText项目下面data文件夹下的models文件夹中的char\_freq.cp文件。其中char\_freq.cp表示SynthText项目下面data文件夹中的newsgroup文件夹下的newsgroup.txt数据集中每个字符的频率。该字符模型在SynthText项目图片合成中作为字符渲染过程中字符选择的以判断标准。

1. 算法简介
2. 输入

词语和词语之间以空格分割开的文本文件，文件编码格式为utf-8（仿照SynthText-master\data\newsgroup\newsgroup.txt文件）。

输入数据格式如下（只截取一小段内容）：



1. 算法

本算法适用于中文，英文文本数据。

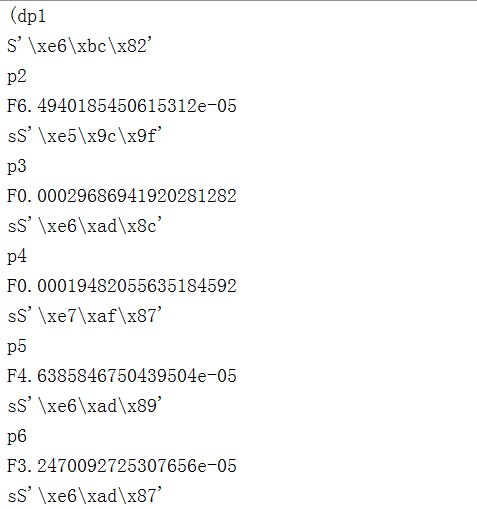
[1]分别统计输入数据中所有字符的出现次数和整个输入数据中总字符数

[2]使用[1]中得到结果，分别计算输入数据中所有字符出现的字频。计算公式为：X字符的字频 = X字符在指定数据集中出现的次数/该指定数据集中的总字符数

(3)输出

将（2）步骤中[2]部分的结果保存到char\_freq.cp文件中，以utf-8编码格式编码。

输出数据格式如下（只截取一小段）：



三、程序实现过程遇到的问题

中文乱码解决办法：

Py文件头部声明# \_\_\*\_\_ coding:utf-8 \_\_\*\_\_

import sys

reload(sys)

sys.setdefaultencoding('utf8')

Python2.7中编码与解码

decode的作用是将其他编码的字符串转换成unicode编码，如str1.decode('gb2312')，表示将gb2312编码的字符串转换成unicode编码。

encode的作用是将unicode编码转换成其他编码的字符串，如str2.encode('gb2312')，表示将unicode编码的字符串转换成gb2312编码。

四、程序运行环境

程序运行环境：Windows7平台，Python2.7.13版本，开发IDE使用pycharm。

运行方法：1、将对应路径修改为你自己的路径 2、选择pycharm中执行命令按钮即可在对应路径生成预处理数据集和相应的字符模型文件。

参考文献:

<http://www.cnblogs.com/fnng/p/5008884.html>

<http://www.cnblogs.com/work115/p/5924446.html>