《特定格式文件内容解析与统计》作业说明

《程序设计实践》作业 2.1

2021 年 06 月

一、 作业内容

编写一个 Python 脚本文件,使用此 Python 脚本,可以读取一个文件夹下所有的 VOC 格式的 XML 数据标注文件的数据,经过解析和统计分析之后,将结果数据存储到一个 JSON 格式的数据标注文件中。

二、 格式说明

2.1 提交方式

在给定的代码文件 `util.py` 中填充指定方法的具体逻辑,实现功能。

最后将`util.py`提交即可。

2.2 命令参数

接口应从参数指定的路径中读取输入目录及输出目录信息,具体参数格式为:

voc_dir:必需参数,指定存放所有 VOC 格式 XML 文件的文件夹的路径

json_path:必需参数,指定存储统计结果数据的 JSON 文件的文件路径

接口应将 json 格式的结果返回，并且将它输出到 json_path 指定的文件中。

2.3 VOC 文件格式

VOC 数据格式详细说明可以参看 VOC 官网,一个样例 VOC 格式的 XML 文件如下:

<annotation>

```
  <folder>VOC2012</folder>
  <filename>2007_000392.jpg</filename>
   //文件名
  <source> //图像来源(不重要)
    <database>The VOC2007 Database</database>
    <annotation>PASCAL VOC2007</annotation>
    <image>flickr</image>
  </source>
  <size> //图像尺寸(长宽以及通道数)
    <width>500</width>
    <height>332</height>
    <depth>3</depth>
  </size>
  <segmented>1</segmented> //是否用于分割(在图像物体识别中  01  无所谓)
  <object> //检测到的物体
    <name>horse</name> //物体类别
    <pose>Right</pose> //拍摄角度
    <truncated>0</truncated> //是否被截断(0  表示完整)
    <difficult>0</difficult> //目标是否难以识别(0  表示容易识别)
    <bndbox> //bounding-box (包含左下角和右上角  xy  坐标)
      <xmin>100</xmin>
      <ymin>96</ymin>
      <xmax>355</xmax>
      <ymax>324</ymax>
    </bndbox>
  </object>
  <object> //检测到多个物体
    <name>person</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>198</xmin>
      <ymin>58</ymin>
      <xmax>286</xmax>
      <ymax>197</ymax>
    </bndbox>
  </object>
</annotation>
```

注:右侧的//是方便理解字段含义手动添加的解释说明,真实的 VOC 格式文件中并不存在。

对于 XML 文件的解析，建议使用相关包（如 xml.dom、xml.sax 或 xml.etree）提

取相关信息。

2.4 JSON 结果文件

对于原始数据,我们需要关注的字段包括:

文件名(filename)

图片尺寸(size)

目标物体的部分标注信息(object 中的 name 和 bndbox 两项信息)

对所有的 VOC 标注数据进行解析之后,我们需要对其进行一定的统计分析,并计算出相应的结果存储到指定的 JSON 文件中,需要给出的字段包括:

标注文件总数量(filenum)

被标注的目标总数量(objectnum)

面积最大的目标信息(maxobject)

面积最小的目标信息(minobject)

为方便理解,假设某个文件夹下只有两个 VOC 格式文件,filename 分别为 2007_000392.jpg 和 2007_000393.jpg,其他内容均和 2.3 中给出的样例内容一致,则此文件夹解析后输出的 JSON 文件内容如样例 JSON 文件所示。

注:样例 VOC 文件和样例 JSON 文件均附在作业说明附件中的 SampleData 文件夹里

2.5 转换说明

请同学们认真查看 2.4 中给出的样例 JSON 格式,最终输出的 JSON 文件格式必须完全一致,这里的一致指的是:1. 每个字段名必须完全一致,包括大小写

也必须完全一致; 2. 每个字段对应的 value 类型,例如字符串类型、数字类型、数组类型和 dict 类型,都必须完全和样例中给出的一致

分析数据中,关于最大最小最多最少的分析,如果发现最值项有多个,则需要把所有的最值项都输出出来,样例中就是这样一种情况,两个最值项都不唯一,请参考样例格式输出。VOC 文件中所有的原始数据字段都存在,不存在字段缺失的情况。

2.6 语法要求

作业提交内容为单个、可正确执行的、符合 Python3 语法的 Python 代码文件。 此文件必须可以在 Python 3.7、3.8、3.9 中的某一稳定子版本上可以正常执行,文件中需要标注使用的 Python 版本。 如果有使用非 Python 自带的第三方的包请在文档中注明。