曲谱分级标题识别流程——草稿版

算法组：郭辉铭

"""

根据位置判定是否标题

难点：

1.存在页眉干扰（去除空格以后单个字体占用的面积来计算 ）

2.标题是否都在曲谱居首的位置 （其实不在）

3.标题是否靠边

4.曲谱内容中存在大段文字

5.标题中有并列，有不同语言的重复的情况

存在居中的标题：

1.先查看位置：之前的标题存在位置居中者（有些页眉的干扰导致无法存在居中者）

2.在y坐标值相差不大且居中的位置继续出现了检测框。则依然认为是标题

3.同一行居中位置，与前一个x位置相差不远的检测框依然认为是标题

4.接下来检测框，与前面的检测框x,y均相差较远，且不再居中的位置，则不再称为标题

不存在居中的标题：

1.依然是主要看第一个检测框，靠近左右的位置

总体原则：先按照这个思路写，写了之后修改

因此大致流程：

step1.首先整合信息至一个列表中并按照单位面积进行排序

step2.取单位面积最大的检测框为首个参照系，判断面积最大的检测框所在的位置：如果在居中位置（这个居中位置不仅包括页面上方，同时包括页面中间位置，

当然存在居中位置，且字体最大者不会是标题，但是目前暂时忽略此种情况），基本可以断定为标题

step3.依次循环遍历多个检测边框（这些边框按照），检测所在位置。如果和第一个位置相当，即都在居中，或者与当前居中者属于x轴上，则认为是标题内容

step4.如果第一个面积最大的边框不在中间位置（）

1:表示标题候选

1.1：表示标题候选的同行

1.2：标题候选的上下

flag的设置：

1.首个flag自然是单位字体最大者

2.接下来则看书否存在同行。若存在则新的flage=[左左上，右右上，右右下，左左下]

同行的定义：

1.满足有轴左边部分在一定的差值范围内

2.两者x轴的差距也在一定的范围内

3.相邻x轴坐标之间的差大于该同行者与边缘的差（包括两边边缘）因此，也要先判定到底在其左边还是在其右边。当然，我已经按照这个区分了。

因此判定同行的流程

step1:从当前节点的左边开始，先判定flag左上和i框右上y的坐标值的差值是否再一定范围内（以0.25*i的边框高度为阈值）

step2:右；计算两者的间隔，并与两者的边缘所在值比较。需要间隔同时<两个边框距离边缘的边缘值

step3:同时满足以上两点的i，则可以称为与flag是同行，但是不改变flag的值

step4：同行绑定，作为上下同列者的判定条件之一

同列标题的定义（一般认为是同样标题的不同意思）

step1:先判定两者之间y轴间隔的值是否在一定值域内（差值小于当前边框i的1.5倍）。只凭借此，则很容易与词曲作者或者序号相混淆

step2:直接检测该边框是否与上一个flag所在的边框的x均值相差不多（主要在于检测误差）：

（但是检测到x均值相差不大的情况下多为中间，假如标题在两边，则需要另外的判断，暂时不做考虑）

若存在：本列检测结束，并记录改序号为标题之一，本行其他内容不是标题。直接开启下一列的判定

若不存在：继续检测同行，直到同行结束

step3:判定本列的同行中哪些是标题选项

1.若只有一个边框，则去除，直接break。当前标题判定完毕

2.若有两个边框，则比较其中间差距与两个边框边缘差的值与两个x的均值。

若x均值类似，但是两个边框相差过大，则不认为是标题。否则，两个均是表达

若x均值不类似，且不存在一个与上个均值类似的边框，则本行不存在标题选项。若存在，则该个为标题

3.若有三个边框，则先检查两个之间的间距：

如果两个边框之间的距离较大，大于其中一个的边缘，则该个排除其中

则接下来继续两个标题的判定

如果三个全部满足边缘距离问题

则三个均为标题候选项

边框间距与边缘值

综：

标题判定选择的流程：

step1: 计算单位字体面积最大者，作为首个flag标记，并测定当前的同行边框，并更新flag标记边框

step2: 左右分别循环（同行边框的首尾）

step3: 先判定是否为同列，如果是同列

step4: 则继续往下找出当前边框的所有同行边框，并更新flag

step5: 判定当前的同行边框是否是标题。若存在标题，则更新flag，若不存在，则本段结束

step5: 如果非同列，则本次循环break，本轮标题查找结束

测试：

分级测试：没步骤完成则测试该步骤

STEP1：测试单位字母面积最大的同行

1.仅一个边框

2.同行两个边框，并列中间，且字体大小相同一份，字体不同一份

3.同行三个边框，一个靠近边缘：左右边缘各一

4.设计两行检测以排除同列错误

step2:测试最大之后的上行标题。分类情况如step1

目前还剩下三个问题：

1.最大字体不是标题者。即，本页无有标题者

2.去除关键字双保险者

3.章节标题不算者

2022.10.20

标题不存在的情况

1.只有章节标题，一般是关键字内：

包含中文，不予检测

不含中文：则统一大小写

仅含数字，乐器名字，仅一个字符，pp，ff， mf，sf，fff

包含速度节拍标记

2.没有章节标题，只有一些关键术语：这种情况下，关键字可以排除掉节拍的影响，但是无法排除掉内部的关键字

最大边框所在的位置：（居中位置或者边缘位置的时候可以认为是标题，如果不是这两项，则不认为是标题）

怎么确定边缘？（边缘最小值在一定范围内）

3.直线检测：查到最大边框周围的直线检测（暂时不做添加）

4.最大边框的首个字符是数字的（暂时不做添加）

因此流程如下：

设定标题与否标志位isExistTitle（=0则本次不存在标题，=1则存在标题）

step1.找到单位字体面积最大者

step2.根据标题所在位置：是否居中或者边缘位置在一定范围内（这个边缘需要考虑，不能太边缘，）

step3.如果2的判定是。则则执行以下.f若为否，则直接退出，输出，该页不存在标题isExistTitle=0

step4.识别该边框。如果包含中文，则执行标题isExistTitle=1

step5.如果不含有中文，则筛查是否为数字，若是则isExistTitle=0

step6.继续检查其中是否只有beatFlag2内的字符串，若是则若是则isExistTitle=0

step7.继续检查，按空格分隔后的字符是否含有beatFlag中出现的字符串，若是则isExistTitle=0

step7.执行完以上步骤，若isExistTitle=1.则执行标题判定。

step8.存储标题边框。（区分字体最大，上同列，下同列，以及同行者）

step9.识别标题，获取接口

串接流程：设定标题存储容器vector<int> title_index;

step1:对整个图片进行边框检测。获取所有边框.然后进行边框排序

step2:循环所有边框：找到单位字体面积最大的这个边框，返回其边框序号

step3:判定边框所在的位置是否居中或者其边缘在一定范围内。若为否，直接退出，不做任何处理

step4:如果step3判定为是。则调用infer_rec识别当前标题候选边框。并判定是否为数字，或者关键字判定。若是,则直接返回。若否，则表示当前候选框认定为标题

step5:执行标题判定。获取标题选项title_index

step6:循环标题选项title_index。识别该边框，获取标题

"""