EDU

CSDN学院　IT实战派

字体反爬

讲师：黄老师

本节目标

1. 学会字体反爬的原理。
2. 学会如何解决字体反爬的问题。

字体反爬原理

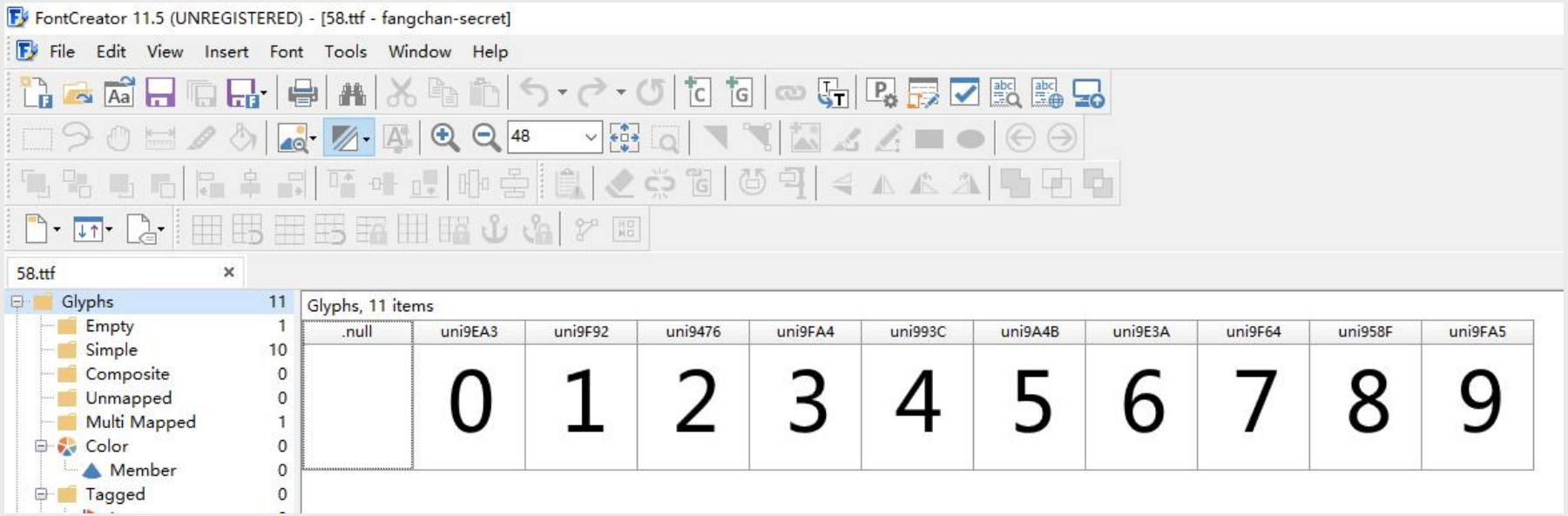
1. 网页开发者自己创造一种字体，因为在字体中每个文字都有其代号，那么以后在网页中不会直接显示这个文字的最终的效果，而是显示他的代号，因此即使获取到了网页中的文本内容，也只是获取到文字的代号，而不是文字本身。
2. 因为创造字体费时费力，并且如果把中国3000多常用汉字都实现，那么这个字体将达到几十兆，也会影响网页的加载。一般情况下为了反爬虫，仅会针对0-9以及少数汉字进行自己单独创建，其他的还是使用用户系统中自带的字体。

字体反爬解决方法-寻找字体

1. 一般情况下为了考虑网页渲染性能，通常网页开发者会把字体编码成base64的方式，因此我们可以到网页中找到@font-face属性，然后获取里面的base64代码，再用Python代码进行解码，然后再保存本地。示例：view-source:https://www.shixiseng.com/intern/inn_a7xabqqr4f9u
2. 如果没有使用base64，还有另外一种方式，就是直接把字体文件放到服务器上，然后前端通过@font-face中的url函数进行加载。示例：https://developer.mozilla.org/zh-CN/docs/Web/CSS/@font-face

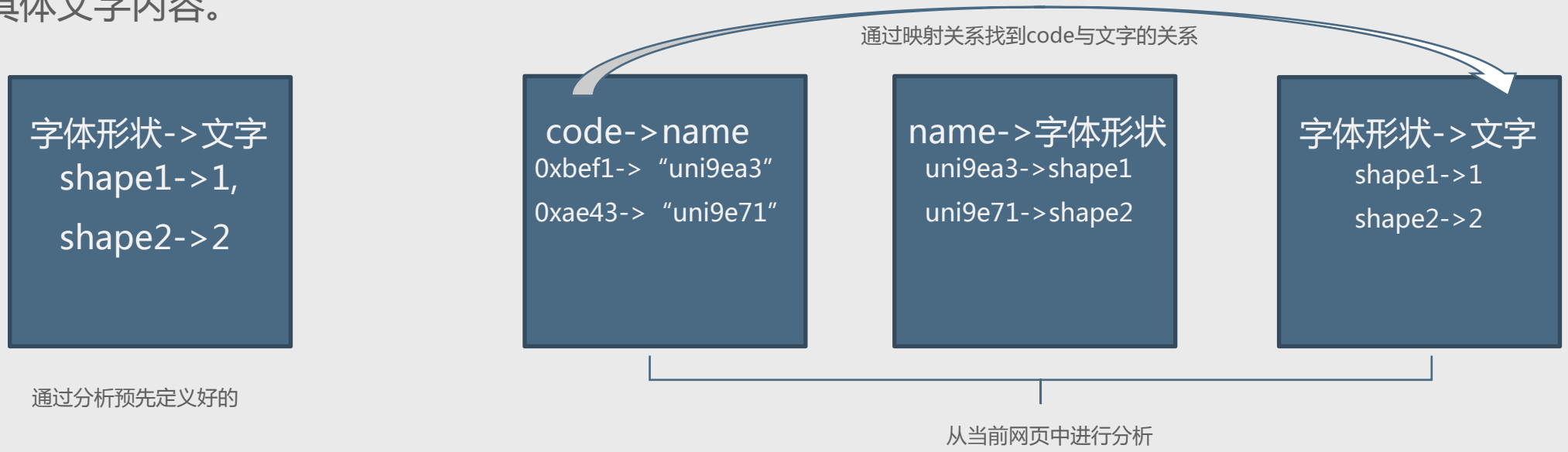
字体反爬解决方法-分析字体

1. 分析字体需要将字体转换成xml文件，然后查看其中的cmap和glyf中的属性。其中cmap存储的是code和name的映射，而glyf下存储的是每个name下的字体绘制规则。
2. 从第1步中我们知道了name对应的字体的绘制规则，但是还是不知道字体是长什么样子，那么可以通过一款叫做FontCreator的软件来打开.tff的字体文件，这样就可以看到每个name对应的字体最终的呈现效果。（FontCreator是一款制作字体的工具，下载地址：https://www.high-logic.com/FontCreatorSetup-x64.exe 这款软件有30天的试用期）。



字体反爬解决方法-根据映射关系获取真实内容

1. 在网页中，直接显示的是字体的code，而不是name。并且网页开发者为了增加爬虫的难度，有可能在多次请求之间code->name->最终字体的映射会发生改变。但是最终字体的形状是不会改变的，因此我们可以通过形状对比来进行判断。
2. 我们可以通过分析字体，得出每个字体形状对应的文字，然后保存到一个字典中。以后再请求网页的时候，就进行反向解析，先获取字体的形状，再通过字体形状反向获取代号所对应的具体文字内容。



58同城字体反爬实战

1. 网址：https://cs.58.com/chuzu/
2. 反爬字体：几室几厅中的数字部分。
3. 字体位置：在网页源代码的@font-face中。

作业-实习僧字体反爬

1. 网址：https://www.shixiseng.com/intern/inn_a7xabqqr4f9u
2. 反爬字体：薪资部分。
3. 字体位置：在网页源代码的@font-face中。

EDU

CSDN学院　IT实战派