电类工程导论(C类)期末大作业项目报告

——表情包检索引擎FaceAlbum

目录

[电类工程导论(C类)期末大作业项目报告 - 1 -](#_Toc503397856)

[1. 项目概述 - 1 -](#_Toc503397857)

[2. 项目环境 - 1 -](#_Toc503397858)

[3. 小组分工 - 1 -](#_Toc503397859)

[4. 项目实现 - 2 -](#_Toc503397860)

[4.1 爬虫获取网页 - 2 -](#_Toc503397861)

[4.2 文字检索 - 2 -](#_Toc503397862)

[4.3 图片检索 - 2 -](#_Toc503397863)

[4.3.1 OCR - 2 -](#_Toc503397864)

[4.3.2 基于SVM和图中文字检索 - 2 -](#_Toc503397865)

[4.4 网页设计 - 3 -](#_Toc503397866)

[5. 项目结果 - 3 -](#_Toc503397867)

[6. 总结与感想 - 3 -](#_Toc503397868)

# 1. 项目概述

近年来，表情包已经成了我们网络聊天过程中必不可少的重要的一环。有时，别人发了一个表情包，我们想发一个别的表情包来回复他（斗图），这时候我们希望以关键字来找表情包；有时，看到一个有趣的表情包，我们想要找到一些与其类似的表情包作为一个系列收藏。

为了解决上述场景中遇到的问题，我们制作了表情包检索引擎FaceAlbum，使得用户可以通过关键字或者图片来搜索表情包。

# 2. 项目环境

Ubuntu14.04 + python2.7 + lucene 4.9.0 + web.py + beautifulsoup + opencv

+ 百度OCR + scikit-learn + scipy

# 3. 小组分工

组长：

贾萧松（516030910548）——OCR(百度API)，拼写检查，SVM标签+OCR文字加权处理，前后端接口实现，图像检索算法设计

组员：

李晋元——图像特征匹配（BOW），SVM分类器（用于获取图片标签），前后端接口实现，图像检索算法设计

曹恒魁——爬虫制作，图片标签索引建立

李明泽——页面设计

# 4. 项目实现

## 4.1 爬虫获取网页

## 4.2 文字检索

这一部分，由于前面的爬虫工作已经获取了每张图片拥有的标签，所以只需要按照中期整合时Lucene索引的建立与检索方式即可。

除此之外，我们还实现了文字检索的智能纠错功能，其主要思路就模糊匹配。

模糊匹配的关键在于如何衡量两个长得很像的单词（或字符串）之间的“差异”。这种差异通常又称为“距离”。

在本项目中，我们使用了基于N-Gram（N元模型）的距离定义，并调用Lucene的suggest模块中SpellChecker实现了该功能，其主要原理就是构建词条字典，检索时找出：与用户输入文本距离小于阈值的字典词条。

其中，词条字典直接来源于已获取的标签索引，能这样做的原因在于我们已经获取的是图片标签（相当于切好的词），而无需对文本分词。

## 4.3 图片检索

### 4.3.1 OCR

本项目中OCR直接使用的百度API，可以获得图片中的文字及其位置（矩形框的四个端点坐标），获得的文字可供后续检索用，而位置可以让后面图像特征提取时不识别文字所在区域，可以减少误判。

### 4.3.2 基于SVM和图中文字检索

利用训练好的SVM分类器，可以从用户给的图片中得到一个最可能的标签，再加上获取到的文字，可以调用Lucene文字索引进行检索。

实现思路为：在Score时，SVM得到的标签设为MUST，而OCR得到的文字设为SHOULD，这里主要考虑到两点：一是OCR可能有错，二是文字可能比较长不易匹配，所以设为SHOULD，使得检索结果必然有SVM标签，而与图中文本越相近，Rank越大，事实证明这样做得到的结果确实是比较好的。

## 4.4 网页设计

# 5. 项目结果

# 6. 总结与感想

在本次大作业中，我们对这一学期学的知识进行了综合的利用，除此之外又进行了许多拓展，由于我们组所选题材——表情包比较特殊，它没有附带信息，它所有的内容都在图片中，所以我们组的工作重心就主要放在了图片检索上，这一点可能与大多数组不太相同。

在学习的过程中，我们对诸如SIFT,SURF,ORB等局部描述子算法的利弊有所了解，以及HOG这种全局描述特征描述算法，并学习使用了SVM这种朴素的分类器，同时对LSH, BOW这两种图像检索算法有所了解，在这个过程中，我们体会到了做图像识别数据库的困难与挑战所在（在这里不得不敬佩ImageNet）。

同时，做搜索引擎的分工合作，锻炼了我们团队协作能力；以及最后的答辩过程，在两位老师的指导下，我们的工作成果汇报能力的得到了很好的锻炼。

最后，感谢何老师和张老师以及助教们这一学期以来的教导与帮助。

2018.1.11