

News Crucible

詹欣宇 谢熠辰 岳野



主要功能



Part I

爬虫



新闻网站

索引



Elastic Search

相似文本推荐



TF-IDF *

**Term Frequency - Inverse Document Frequency*
是一种用于信息检索与数据挖掘的常用加权技术。TF意思是词频，IDF意思是逆文本频率指数。


1.1 索引

使用Elastic Search 代替Lucene进行检索，更加快速、稳定、可靠；
更加方便得获得我们需要的各部分信息；
更好得兼容python 3，便于使用；

1.2 相似文本推荐

使用TF-IDF对文本内容进行比对；
通过设置停用词的方式避免一些常见词的影响；
如果某个词或短语在一篇文章中出现的频率高，
并且包含这个词条的文档数目较少，那么这个词
或者短语就有很好的区分能力；

Part II

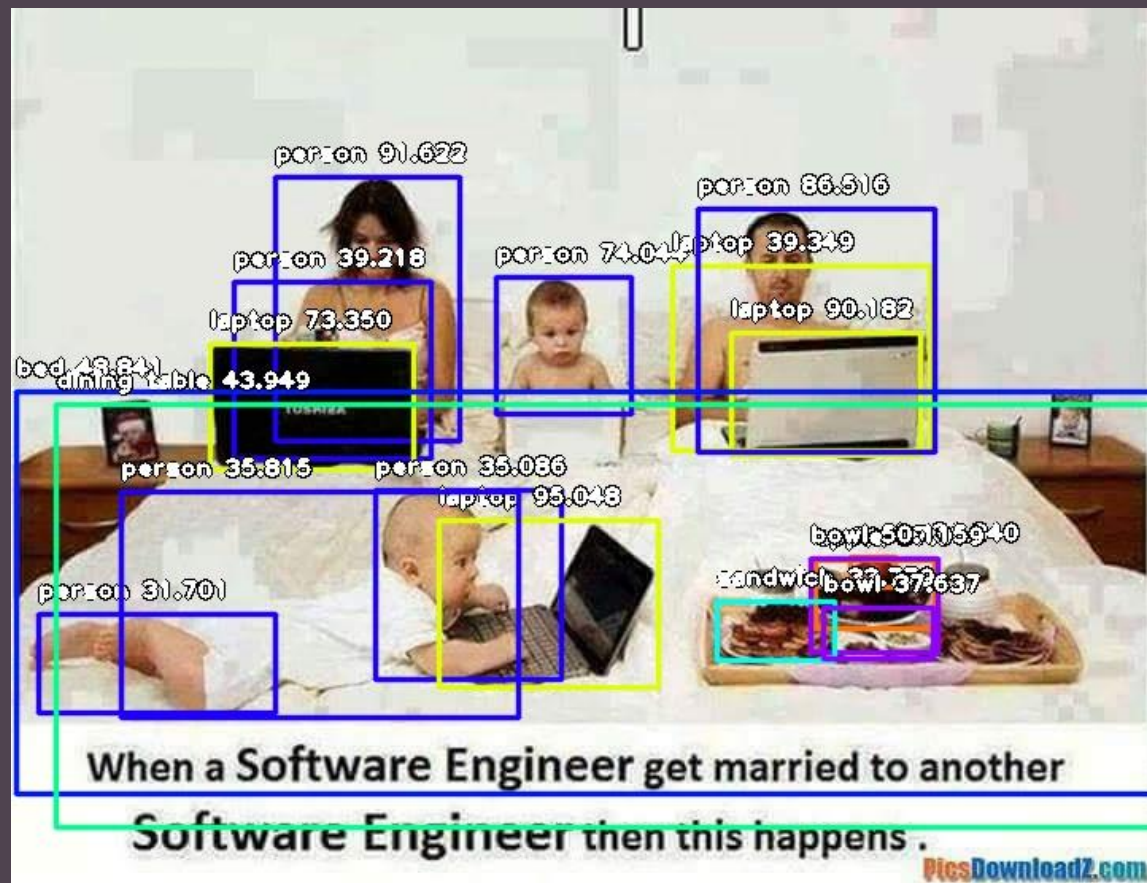
图片搜索文字  深度学习
TensorFlow
提取特征

图片搜索图片  感知哈希 初步筛选
SIFT 提取特征 细致比对

2.1 图片特征提取

使用TensorFlow神经网络，通过深度学习的方式进行目标检测

先识别出图片中的每个目标的位置，再进一步确定目标的内容



2.2 相似图片查找

初筛：

使用感知哈希算法(pHash),采用DCT（离散余弦变换）来降低频率；
通过汉明距离来比对两张图片直接的相似度；
筛选出一小部分，进行细致比对；

细筛：

使用SIFT算法找出图片的一定数量的特征点,通过比对特征点的方式来找出相似图片；
这样可以更好的识别旋转、平移等图片变换；

Part III

Web 框架



tornado

CSS/HTML 框架



Bootstrap 4