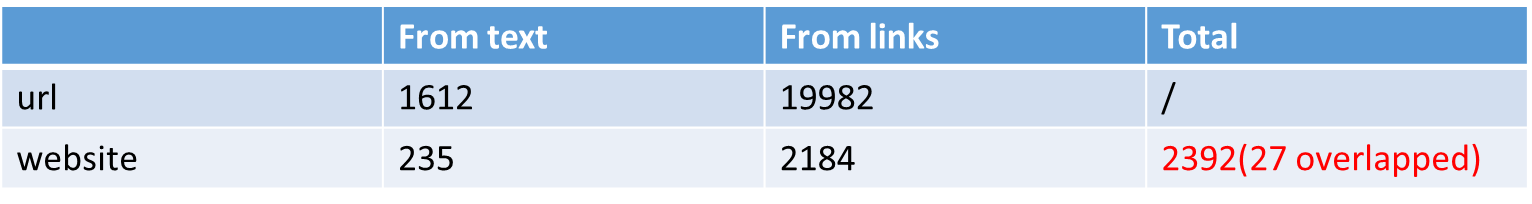
**双周报--智能舆情分析及黑产挖掘相关**

1. **电商黑灰产自动化监测及评估(刘倩君)**

进展：

从舆情文本和论坛等链接提取候选网站2392个待检测，过滤掉国内top100网站和国外常用网站后还剩2303个候选，爬取了这些网站的前100个页面开始进行特征提取。预测这些网站中黑灰产网站比例在10%左右，所以后面还会继续扩充，一个可行的方法是从已检测到的黑产网站中获取出站链接，另一方面需要从舆情文本中提取命名实体。

**候选网站来源：**



**特征提取：**

（1） site name:

e.g.

顺风空包网 [http://www.sfkongbao.com](http://www.sfkongbao.com/)

5v空包网 [http://www.5vkongbao.com](http://www.5vkongbao.com/)

淘优网[http://www.5v186.com](http://www.5v186.com/)

空包流量网[http://www.kb698.com](http://www.kb698.com/)

空包 [http://www.kongke8.com](http://www.kongke8.com/)

The site names have similarities .

（2）page content:

n-gram terms :

e.g. 任务，代刷，信誉；刷流量，不降权；打造爆款，购买空包

html tag as a tuple（type，attribute，content）：

e.g.



This tag implies whether users needs to register and verify themselves to log in. Most black underground websites require users to provide their phone number while most forums and bbs do not.

（3） traffic statics:

page view , IP

(4) links to known black websites

1. **黑词挖掘和扩充（伍一鸣）**

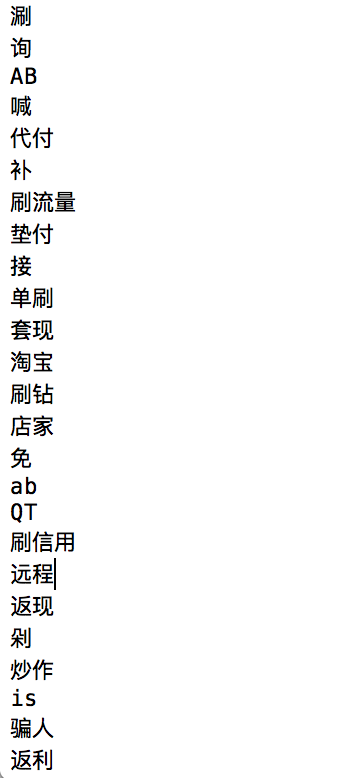
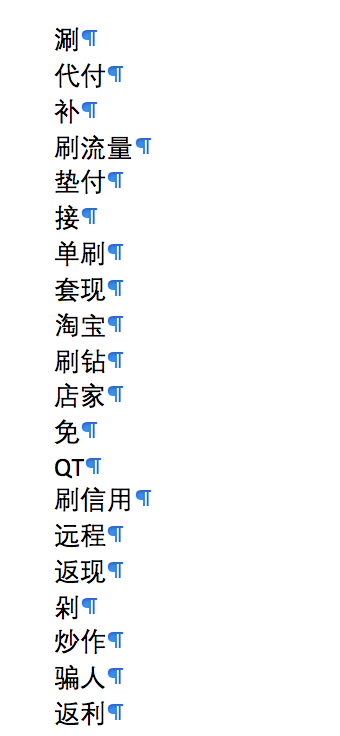
4月份主要是在完善黑词库，主要利用下面两种思路：

idea 1:选取相关性较高的文本 --> 分词 --> 训练word2vec --> 以爬虫关键词为种子从w2v词典中计算与它距离最近的词，取topK，作为种子的相关词—>人工标注出黑白词

idea 2: 抽取关键词—>利用word2vec将语料库中的每个词都转换成词向量—>将关键词的词向量进行聚类

（1）两种思路都已经完成，由于idea1没有对文本进行去停顿次等处理，所有最后的结果涵盖较多的杂词，人工标注的代价太高。并且认为，以种子词作为中心所聚类得到的黑词是具有局限性的，只会得到与该种子词较近的词，而种子词没有涵盖所有黑词的类型，所以认为这种方法会丢失其他类型的词

人工标注前的部分结果： 人工标注后的黑词

（2） 所以我个人比较倾向于idea2，并且已经完整的做完，还存在一些小的问题来提高精确度。

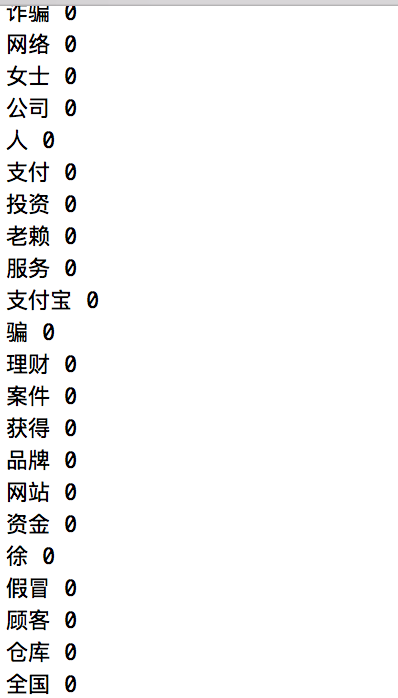
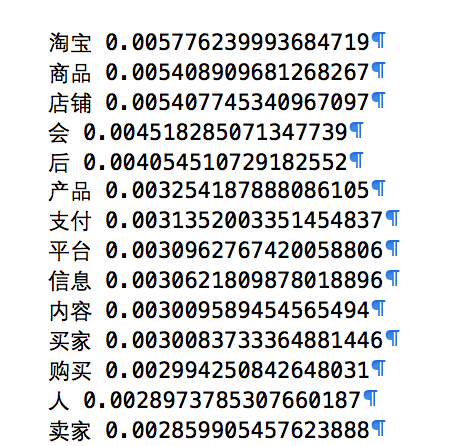
关键词的提取：利用textrank关键词提取算法

过程:分词(tool:jieba分词)标注词性—>去停用—>对词性过滤（留下’an’, ‘i’, ‘j’, ‘l’, ‘n’, ‘nr’, ‘nrfg’, ‘ns’, ‘nt’, ‘nz’, ‘t’, ‘v’, ‘vd’, ‘vn’, ‘eng’ ）—>构造无向图—>迭代运算textrank算法对无向有权图中的词进行降序排序，输出topK个词作为关键词

聚类：利用基于余弦距离的kmeans++算法。

目前基于余弦距离的程序不太稳定需要改进，暂时展现基于欧式距离的聚类

关键词部分结果： 聚类部分结果：



(3)改进:

idea2目前的结果也存在夹杂白词的现象，需改进算法。

目前可以改进的地方：

对词性过滤的程序进行测试，更准确

不断调整关键词的个数，聚类的簇数

使用基于余弦距离的kmeans++算法

(important)使用外部库，例维基百科。在文本中加入外部库，拉大黑词与白词之间的距离

1. **微博言论识别和挖掘（吕晨阳）**

遇到了以下问题：申请各种权限；微博经常强化反爬虫监控，爬虫经常失效经常修改代码，尝试给爬虫加IP代理池，发现效果很差，最终改成分布式爬虫。受限于资源，当前爬取速度大约是三个月40万个节点。

进展：当前分布式爬虫趋于稳定，虽然当前微博再次改变安全策略爬虫失效，不过预计能很快解决。阅读论文思考京东迁移学习模型。正在做毕业设计。