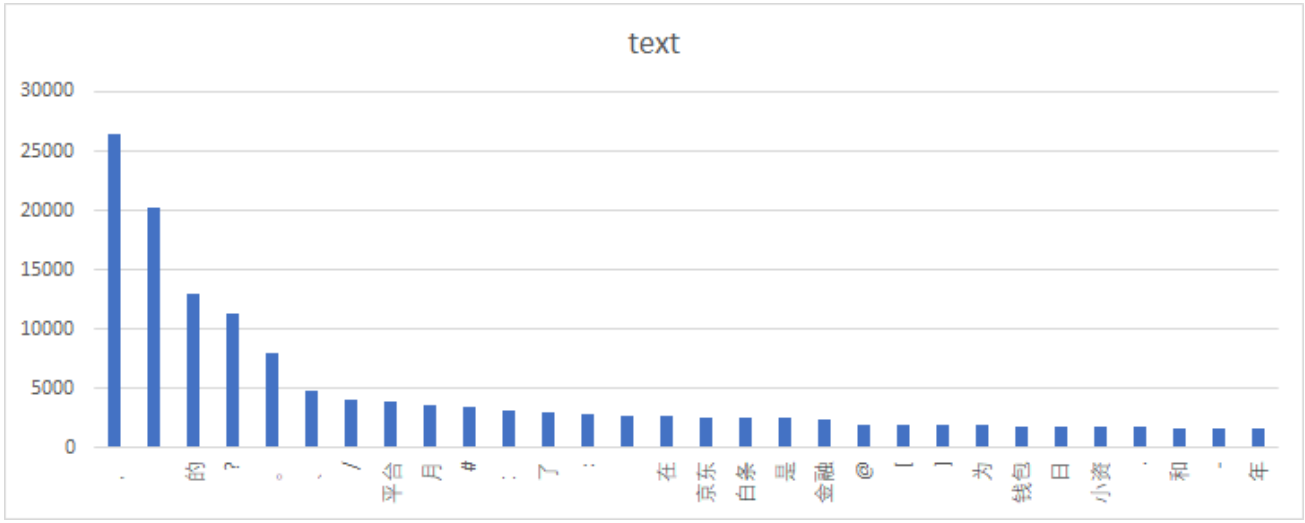
1. 统计第一版分词+词频+tfidf



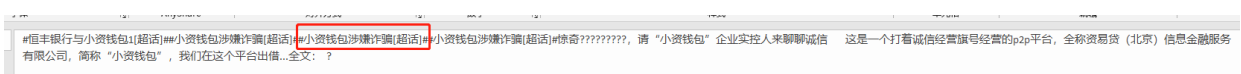

```
['白条', '小资', '京东', '超话', '钱包', '##', '提现', '诈骗', '平台', '出借', '...', '资易贷', 'nan', '海淀', '北京', 'P2P', '全文', '涉嫌', '网贷', '宜贷']
++++++++++++++++++++++++
['白条', '京东', '平台', '小资', '超话', '2018', '提现', '网贷', '钱包', '##', 'IMG', 'P2P', '金融', '诈骗', '06', '出借', '05', 'http', 'cn', '...']
```

根据第一版分词统计，设计正则匹配式

1. < >
2. http
3. 邮箱

   数量很少，也可以不匹配

4. tag



有**可能**过滤掉有效信息or实体

5. 表情{IMG}
6. 年份
7. 空字符

~~正则好过滤一些就没放在停词里~~

2. 通过正则表达式匹配这七项，统计出他们的频率后发现需要过滤，所以开启正则过滤

~~*虚假的推理步骤，其实本来就准备过滤。当然也可以通过这步推出无须过滤邮箱。*~~

~~懒得画图了，感觉后期写报告可以再细画~~

七项匹配次数分别为

585

675

24

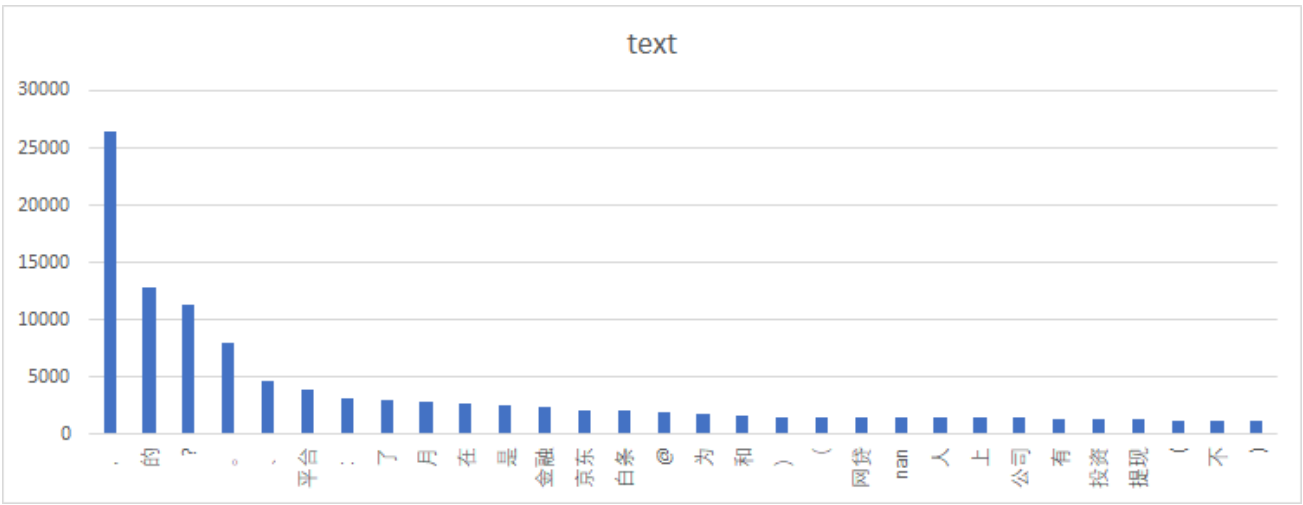
3075

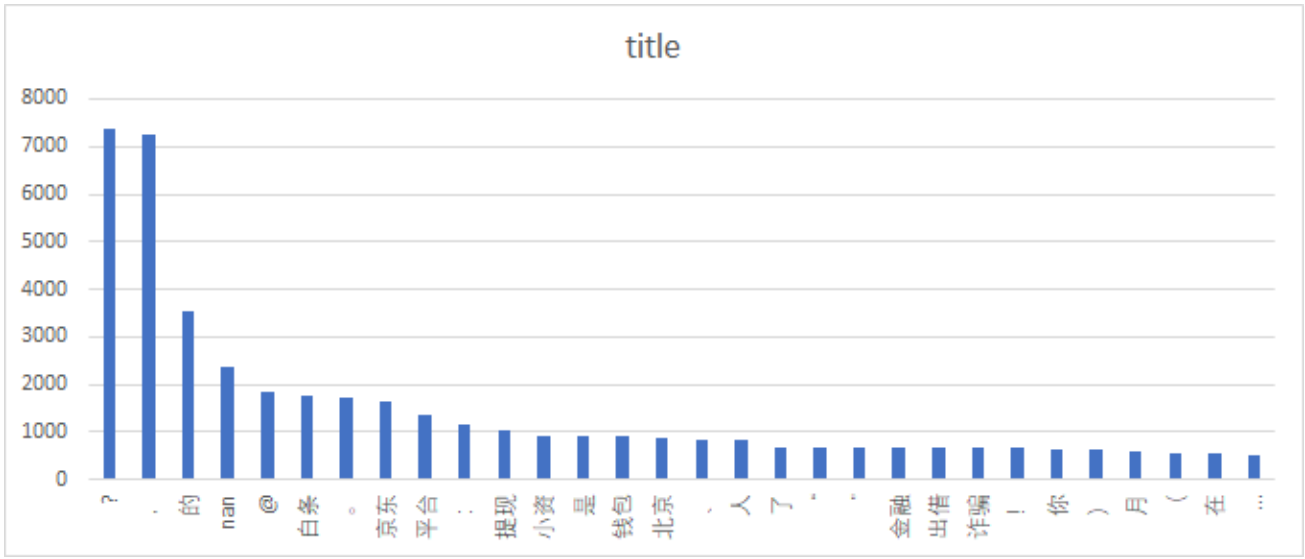
1250

1860

16384

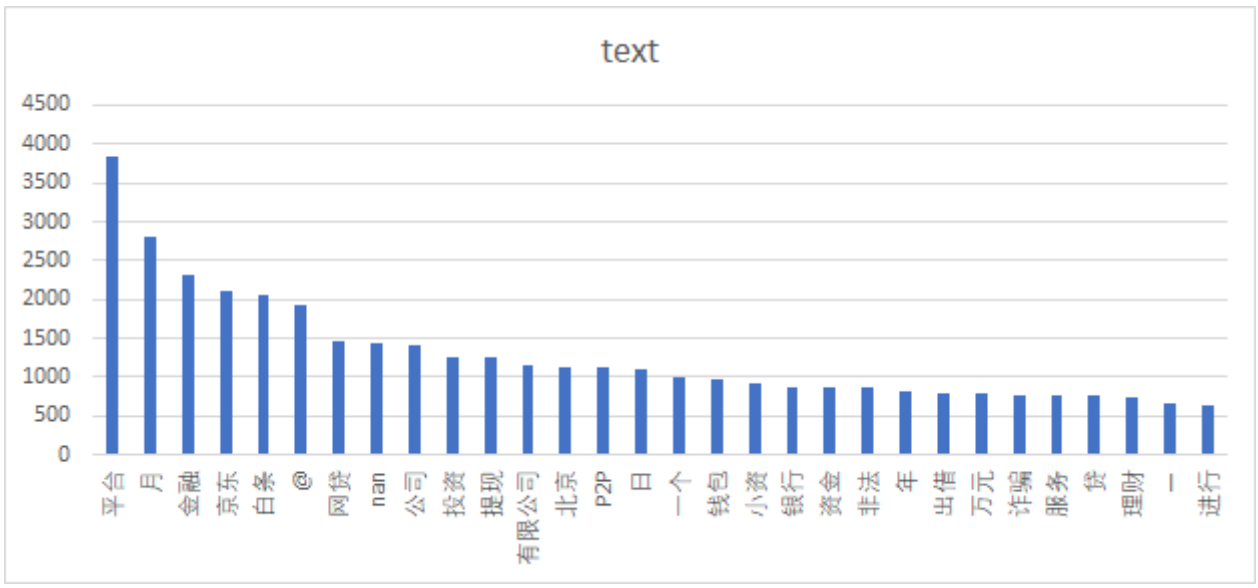
3. 开启正则过滤后再统计一版词频，计算tfidf，统计需要过滤的停词

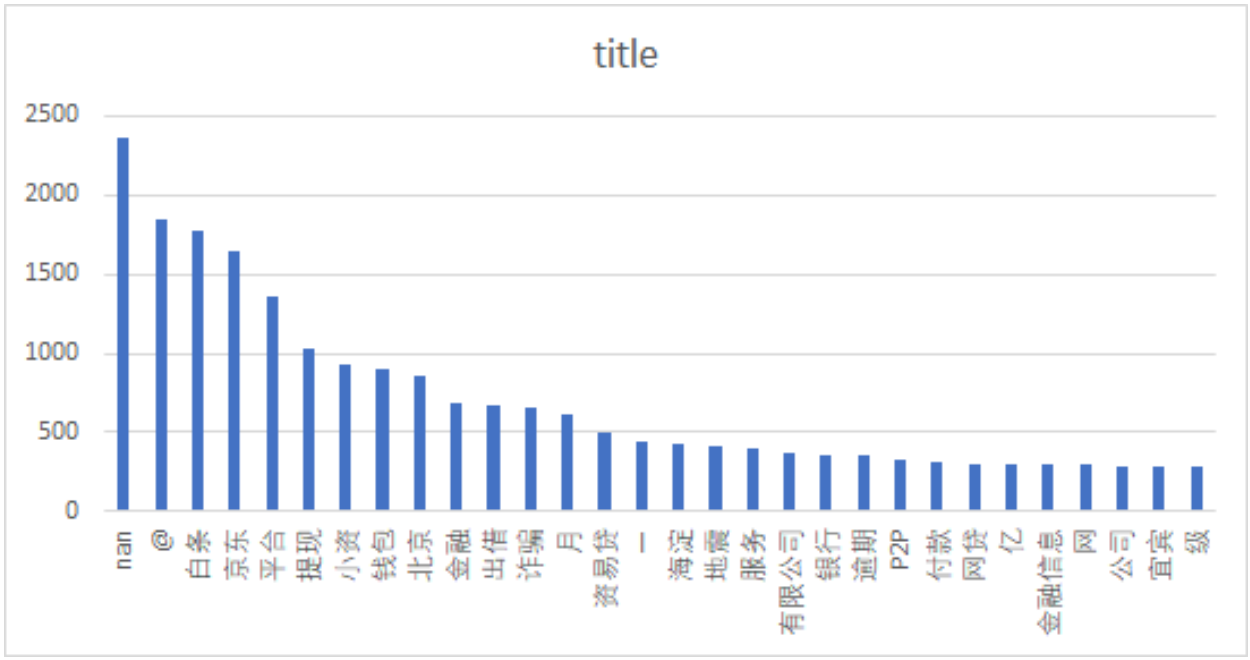





```
['白条', '京东', '提现', '小资', '平台', '钱包', '出借', '...', '诈骗', '资易贷', 'nan', '海淀', '北京', '全文', 'P2P', '网贷', '金融', '逾期', '宜贷', '金融信息']
++++++++++++++++++++++++
['白条', '平台', '京东', '网贷', '提现', 'P2P', '金融', '小资', '钱包', '出借', '...', '诈骗', '非法', '联璧', '有限公司', '资易贷', '逾期', '北京', '公司', '理财']
```

4. 开启停词过滤后的词频统计+tfidf






```
['白条', '京东', '提现', '小资', '平台', '钱包', '出借', '诈骗', '资易贷', 'nan', '海淀', '北京', 'P2P', '网贷', '金融', '逾期', '宜贷', '金融信息', '24', '地震']
++++++++++++++++++++++++
['白条', '平台', '京东', '网贷', '提现', 'P2P', '金融', '小资', '钱包', '出借', '诈骗', '非法', '联璧', '有限公司', '资易贷', '逾期', '北京', '公司', '理财', '投资']
```

5. 最终输出过滤后的文本串到clean.csv中

   ~~*只处理了train.csv，敲定清洗策略后再处理test.csv*~~

Q

记者?