# 配置文件说明

config.json 是整个工程的配置文件，定义了不同的操作流程以及各操作流程中自定义的参数。 流程是由stage 控制的。

## 数据清洗 stage=1

修改config.json中的"stage": 1, 并且把合并后的原始数据放入data文件夹下，并且改名为all\_tweets.txt。 运行脚本， 由于还计算了打分，因此运行时间比较长

## 词频计算及共现矩阵 stage=2

修改config.json中的"stage": 2. 第一步完成后，data文件夹下会生成一个all\_tweets\_clean.txt的文件，可以打开手工观察。文件比较大，可以把前1000行过滤出来, head –n 1000 all\_tweets\_clean.txt >head\_1000.txt. 会在output文件夹下生成两个文件，分别是词频和共现矩阵。

|  |  |  |
| --- | --- | --- |
| use | clean\_text\_cust | 使用清洗后的clean\_text\_cust计算词频  text: 原始的; clean\_text: 简单清理的 |
| ignore |  | 忽略的此列表，停用词 |
| transform |  | 替换 |
| maxWord | 30 | 只关注前30个高频词 |
| frequencyThreshold | 3000 | 只关注出现频次大于3000的词 |

这一步可以反复调试

## 聚类

修改config.json中的"stage": 3. 修改num\_of\_cluster， 最终聚类的类别个数

细节： 第2步完成后得到的是一个相关系数矩阵，需要在此基础上定义距离矩阵。 目前定义方式为 距离= - log(相关系数)。 如果要改变距离定义方式，请联系技术支持。

[聚类系列-层次聚类(Hierarchical Clustering) - CSDN博客](https://www.baidu.com/link?url=7mLOAweA2VvKlgJsFPIb2gfDN-pDJKx2A0QioKhEFw8aN1bQScQ8ybn13AkvEZrh_G1sQNoqvQccoCo6akRdv_q2MQfGuTl48lEZXZ7OvMG&wd=&eqid=cea12a9a0001c62d000000035b0a6d4b)

聚类结果 会输出到命令行。 在output/cluster 有详细的分类之后的tweet数据

## 情感分析可视化

|  |  |  |
| --- | --- | --- |
| sentimentMethod | baseonscore  compound | 定义正性和负性的方法  Baseonscore：需要和neutralThreshold  结合使用 |
| neutralThreshold | 0.66 |  |
| plot\_data\_file |  | 要画图的文件，可以是整个的，也可以是聚类后的 |