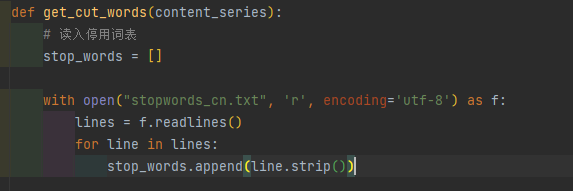
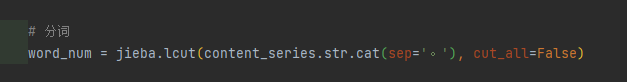
分析报告

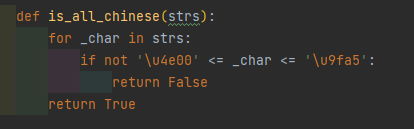
根据数据集，对数据集进行分词，采用的是jieba分词，首先先用停用词去把无效词去掉



接着根据标准模式，对句子进行分词处理

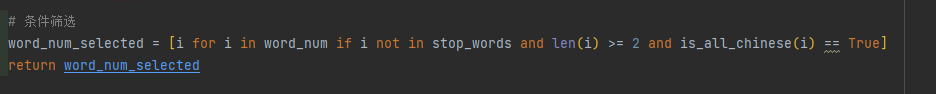


再去进行条件筛选，把存在停用词文本和词数小于2个去掉，保留有用词，再去判断是否为中文



这里采用了正则表达式来对文本进行判断，在\u4e00- \u9fa5范围区间则是中文

不符合则排除掉



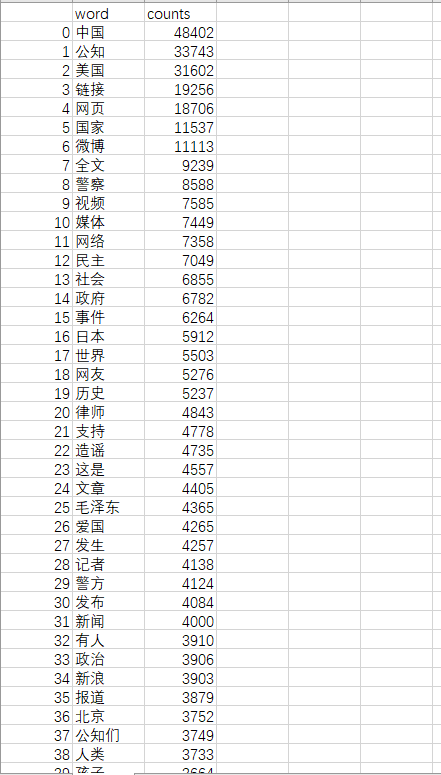
然后返回处理好的分词

接着采用stylecloud进行分词可视化





再去统计词的频次用pandas把词频保存为CSV表



接着我们根据高频词的top来进行词频分类



这里一共有5个类别

原文是讲述公知的有关内容，所以可以大致分为以上四类

社媒类，对应的就是把传播的方式，传播的平台，传媒的文件名，全部分为一个类

正面类，则是在讨论公知时，所涉及我国的一些专有名词，提到中国，国家，社会，政府，毛泽东，爱国，台湾，北京，这些都是我国的一些特有名词，多为正面倾向，

而提到孩子，民族，则是表示公知会对这些目标群体所造成危害性

负面类，则是在阐述公知的由来，由西方国家通过社媒舆论来对我国进行文化侵略的一种途径，大多为负面舆论

事件发展类，对应事件产生的由来，这事件发展的情况，以及借鉴历史事件，希望人们能警惕文化入侵的危害，唤醒人们的意志

职业类，则是各种职业的归类