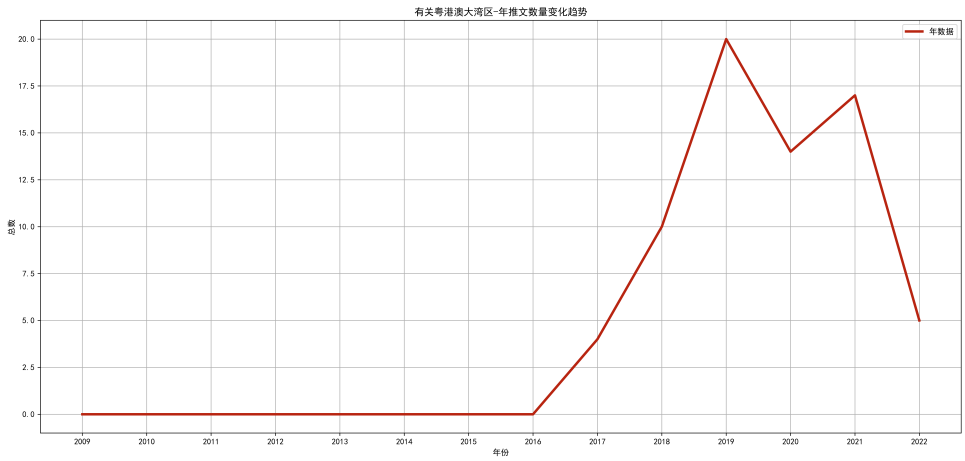
根据以下要求进行数据分析

1、推特中每年有多少推文是关于粤港澳大湾区。



这题的思路：

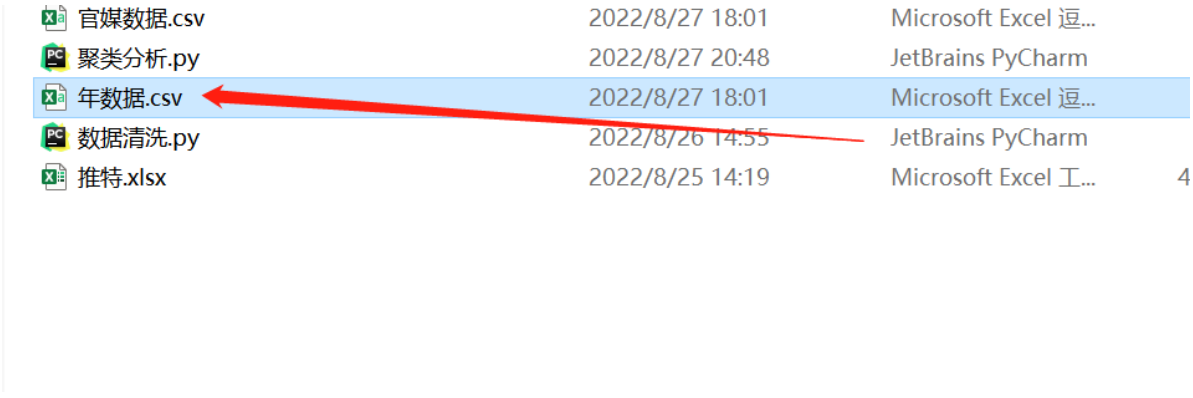
首先想要知道每年有多少推文是关于粤港澳大湾区，首先先判断该推文是否是和粤港澳大湾区有关

这时候我们就去判断，该推文是否有如下关键词：Guangdong、Hong、Macao、Greater Bay Area，这些关键词出现了，就说明是有提到粤港澳大湾区

那么我们就判断该推文是提及到粤港澳大湾区的，然后进行统计，统计好之后，再去统计全年的数据，一年出现了多少次，然后再去用可视化的方式更加

直观的表现出来。

如果想要看具体数据的话，则是查看相关文档，这里面清楚的记载了，每一年的数据



2、多少人关注粤港澳大湾区。

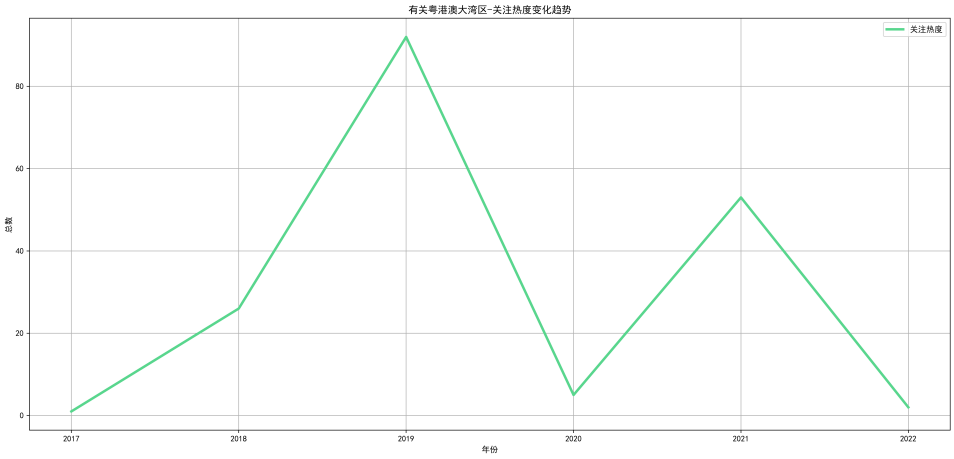
对于多少人关注粤港澳大湾区，这个的话，思路如下：

首先先找出相关推文，这一步就是第一步的内容，因为第一步已经判断好了，哪一些推文是在讨论粤港澳大湾区的，

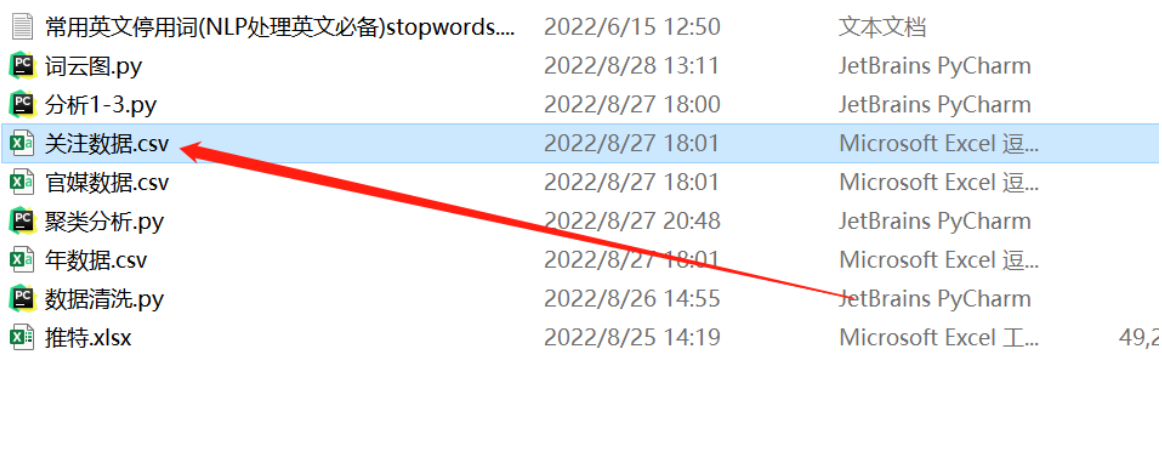
然后我们再根据这些推文，有多少人评论，把这些评论数全部加起来，那么就是可以算出每年有多少人在关注粤港澳大湾区，因为评论是最直观的

，一般点赞，转发并不能很直观的说明，这个人一定在关注粤港澳大湾区的，但是如果已经去评论了，那么这个人一定是关注粤港澳大湾区的

所以我们就是去统计每年相关推文，底下的评论量是多少，然后再用可视化的方式表达出来



相关数据的话，则是去这里查看

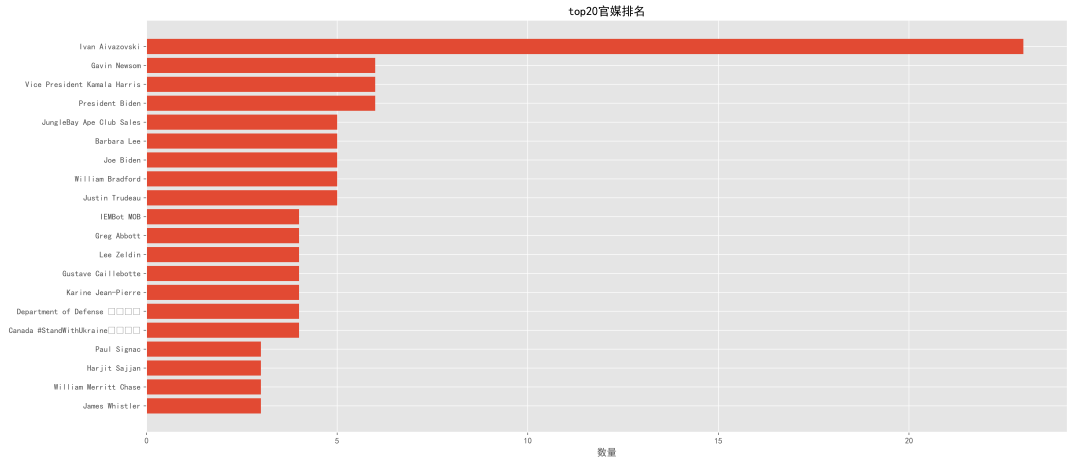


3、有哪些国内外官媒或者自媒体关注粤港澳大湾区。

至于这一点，首先的话，默认所以的数据都是涉及到粤港澳大湾区，不然也不会去收集这样的数据，既然所有文章都是涉及到粤港澳大湾区的

然后我们先把官媒的账户指定出来，

指定出来后，然后去对其进行统计，查看哪一个官号发推文数量最多，就说明这个官号是最活跃的，进行排序，选出前20活跃的官号



这些是相对来说，最活跃的官号，其实整体有300多个官号，只是因为数量太多了，所以选前20的官号就好了

查看具体数据的话，如下

名称	修改日期	类型	大小
data	2022/8/28 12:58	文件夹	
class-fenci.txt	2022/8/27 20:21	文本文档	1,522 KB
img01.png	2022/8/27 18:01	看图王 PNG 图片文件	450 KB
img02.png	2022/8/27 18:01	看图王 PNG 图片文件	514 KB
img03.png	2022/8/27 18:01	看图王 PNG 图片文件	448 KB
new_推特.csv	2022/8/26 15:12	Microsoft Excel 逗...	106,904 KB
常用英文停用词(NLP处理英文必备)stopwords....	2022/6/15 12:50	文本文档	9 KB
词云图.py	2022/8/28 13:11	JetBrains PyCharm	3 KB
分析1-3.py	2022/8/27 18:00	JetBrains PyCharm	3 KB
关注数据.csv	2022/8/27 18:01	Microsoft Excel 逗...	1 KB
官媒数据.csv	2022/8/27 18:01	Microsoft Excel 逗...	5 KB
聚类分析.py	2022/8/27 20:48	JetBrains PyCharm	4 KB
年数据.csv	2022/8/27 18:01	Microsoft Excel 逗...	1 KB
数据清洗.py	2022/8/26 14:55	JetBrains PyCharm	4 KB
推特.xlsx	2022/8/25 14:19	Microsoft Excel 工...	49,268 KB

4、推文的相关内容是什么。

首先根据需求，想要找出7大主题内容，根据这个思路，分为两步

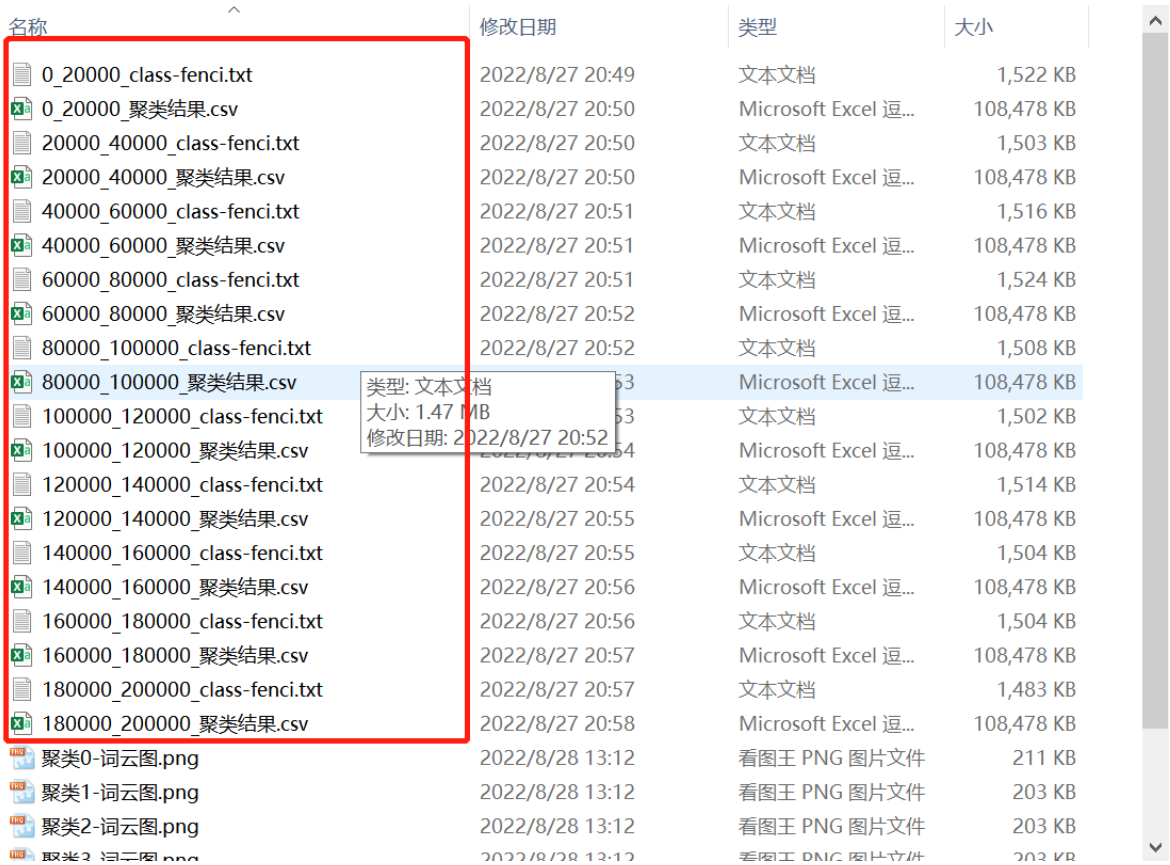
一是人工判断，二是采用机器学习里面的无监督算法学习，k-means，聚类

因为考虑到有20万的数据，所以人工是不现实的，这时候我们就是采用机器学习里面的无监督学习

去对文本进行自动分类，自动打标签，把全部文本分为7大类

因为数据量过大的缘故，所以我们采用分段去进行分类

把20万的数据，划分为10个文本，然后再去进行分类



然后根据分好的类别，我们进行词云图把相关数据展示出来，来查看，这个分类具体是在讲述一个什么样的事情

首先聚类0的词云图如下

这里面的高频词为police,friend,support ,night，county,community,life,storm,weather,local,这些词，哪些词大，就说明该词出现的频率越高

可以看出，这些高频词，其实都是和生活的词有关，那么该主题主要是涉及到民生

其余的聚类，也是按照这样的方法去判断，先查看它的高频词如何，然后通过词云图的方法来表达出来，该聚类是涉及到什么内容

然后对其进行归类，因为内容较为重复，这里就不过多讲述了，你按照我这样的方法去判断，该聚类是关于什么样的主题即可