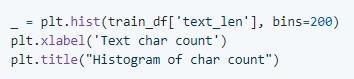
Task 2 : 参照笔记：https://zhuanlan.zhihu.com/p/162749684







绘制直方图：



新闻类别分布：



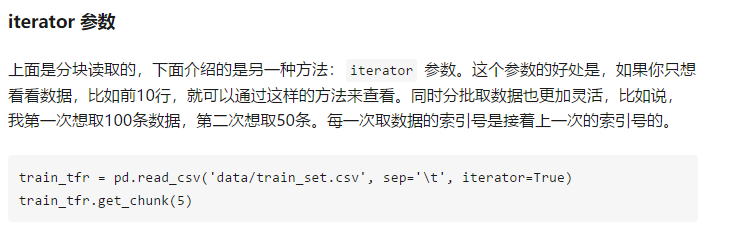
字符统计分布：



编号3750的字出现的最多，编号3133的字出现的最少，总共有6869个字

word\_count = sorted(word\_count.items(), key=lambda d:d[1], reverse = True)

# 对{字符:字频数}列表 按 值 进行排序，也就是按字频数进行排序。

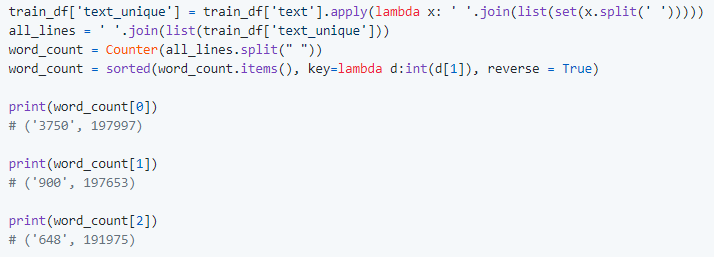


Example：





这里还可以根据字在每个句子的出现情况，反推出标点符号。下面代码统计了不同字符在句子中出现的次数，其中字符3750，字符900和字符648在20w新闻的覆盖率接近99%，很有可能是标点符号。



set() method is used to convert any of the iterable to sequence of iterable elements with distinct elements, commonly called Set.

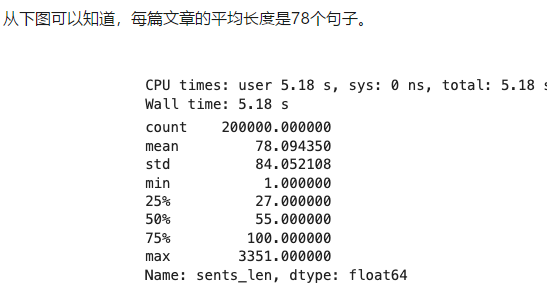
HW1:

* 假设字符3750，字符900和字符648是句子的标点符号，请分析赛题每篇新闻平均由多少个句子构成？

%%time

train\_df['sents\_len'] = train\_df['text'].apply(lambda x: len(re.split(' 3750 | 900 | 648 ', x)))

train\_df['sents\_len'].describe()



HW2:

* 统计每类新闻中出现次数对多的字符
* %%time
* train\_df = pd.read\_csv('data/train\_set.csv', sep='\t')
* train\_df['text'] = train\_df['text'].apply(lambda x: x.replace(' 3750',"").replace(' 900',"").replace(' 648',""))
* # 同一类的拼接到一起
* for i in range(0, 14):
* df = train\_df[train\_df['label'] == i]['text']
* word\_count = Counter(" ".join(df.values.tolist()).split())
* print(i, word\_count.most\_common(1)[0]) # 新闻类，次数最多的字符及次数