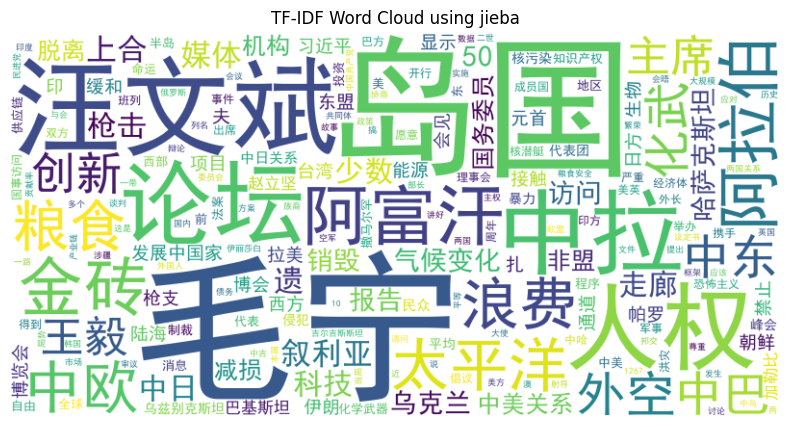
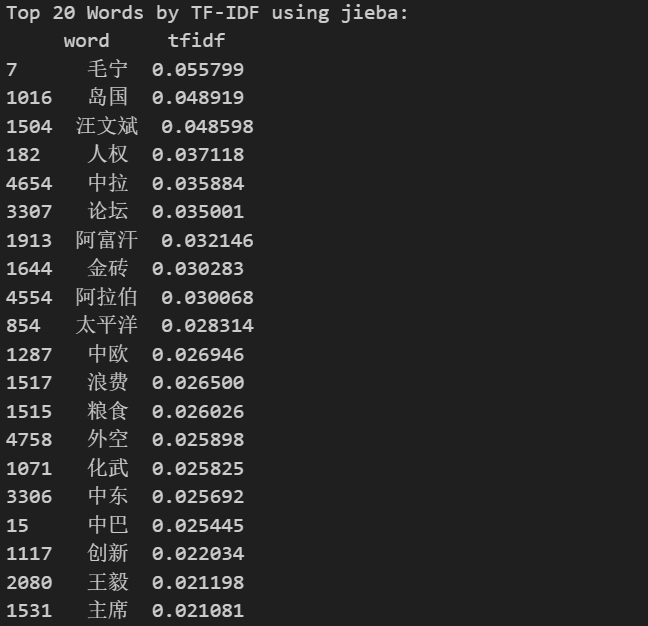
第二次实验报告

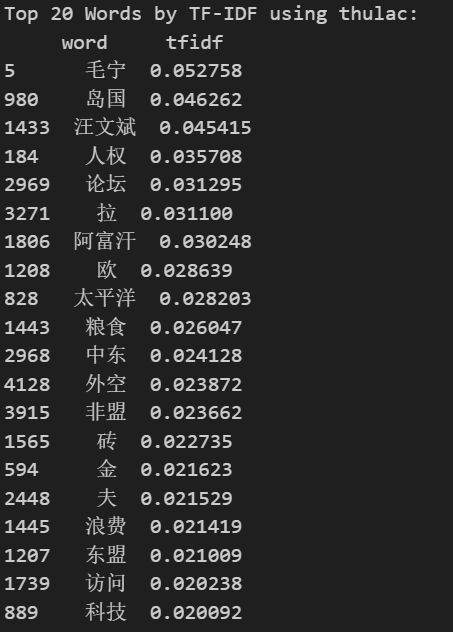
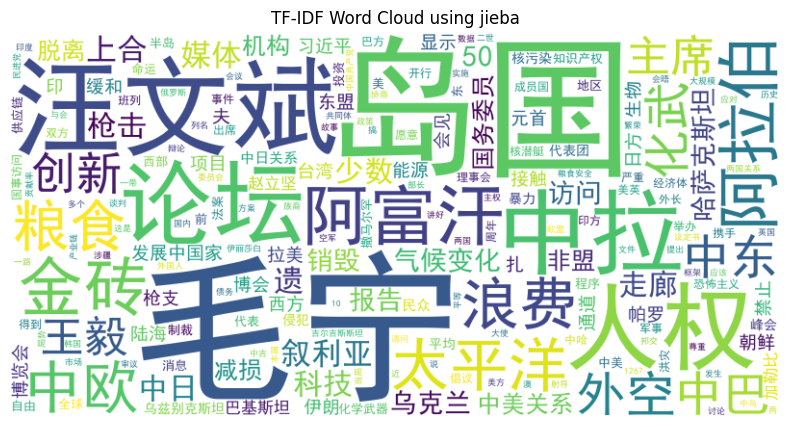
唐屹10235501451

1. 实验结果展示：

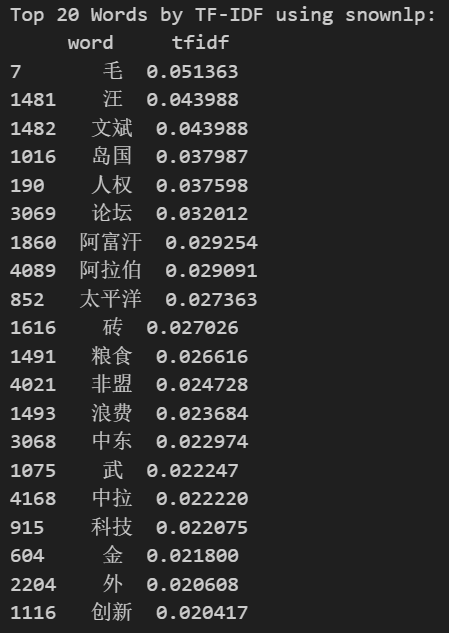
本次实验分别调用四个分词库（jieba、thulac、snownlp、nlpir）进行分词和词云生成，并且输出在不同分词方法下词频前20的单词（创新点）

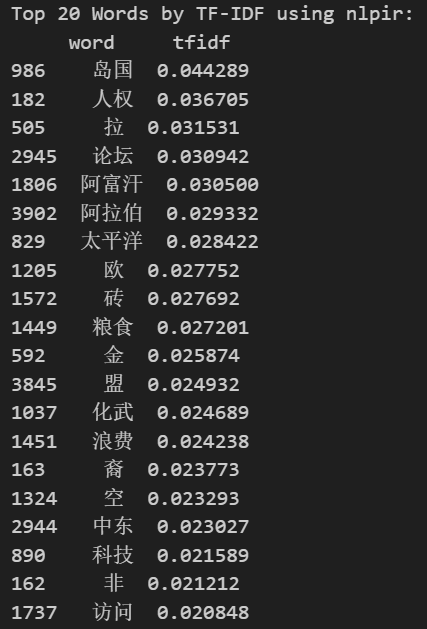
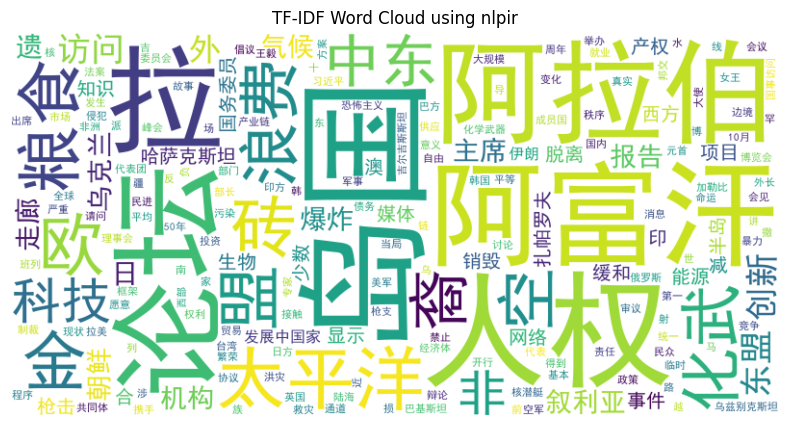
Jieba库分词词云



Thulac分词库词云

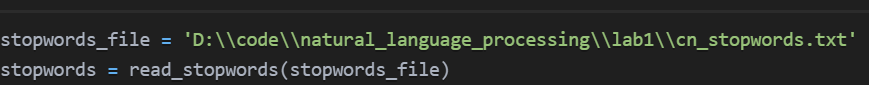
Snownlp分词词云



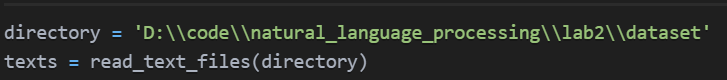
Nlpir分词词云

1. 代码实现：

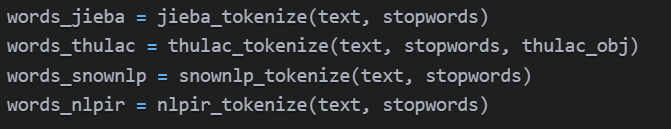
停用词读取



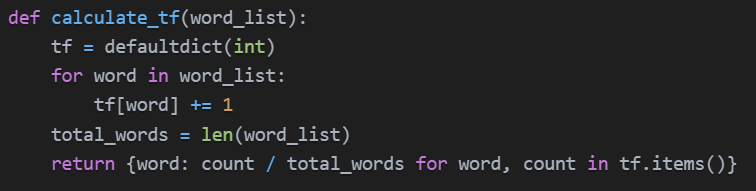
读取文本（用read\_text\_files函数从指定文件夹读取多个文件）



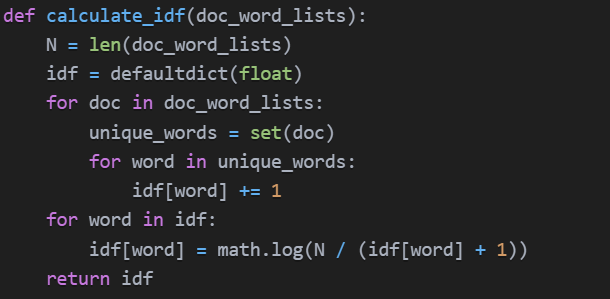
调用分词库



TF计算



IDF计算



完整代码可见本人github仓库[natural\_language\_processing/lab2/lab2.ipynb at main · dingzhen1145141919/natural\_language\_processing](https://github.com/dingzhen1145141919/natural_language_processing/blob/main/lab2/lab2.ipynb)

1. 实验结果分析

由于实验使用了多个分词库，分词展示的词云各不相同，但是某些词库有明显缺陷，比如jieba、thulac库无法将50给识别为禁用词（禁用词词典里“5”、“0”均为禁用词，但组合起来没有被禁用，猜测是分词原理导致），snownlp分词将词分得太过碎片化，如“汪文斌”的“汪”被单独分出来“，nlpir在这方面稍优于前者。根据高频词输出结果来看，对于新闻类的文本来说使用jieba库分词能够较为准确地展示文本主要内容。

Jieba库词云的“50“ thulac词云的”50“ snownlp词云的”汪“