**关键词抽取实验报告**

本次实验采用了上一次已经标注过的人民日报的语料进行分析，使用TF-IDF算法进行关键词抽取，就要建立相应的语料库，对文档内容进行切分。

大体思路有两种：按照关键词切分和按照行数切分。

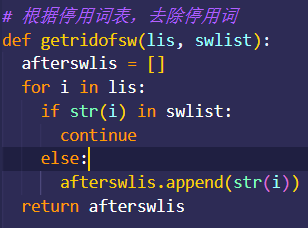
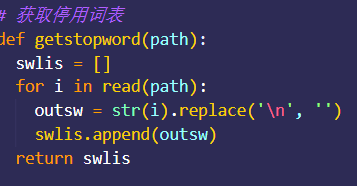
首先采取了按照关键词切分，即使用关键词：本报，经过核验，准确率不高，文档之间大小差距也比较明显。随后采取列数划分的办法，以90行为划分标准，对文档进行划分，最终所得49个切分文档，将所得文档放在input文件夹内，input文件夹作为一个语料库。

1.py为切分文档的python程序。

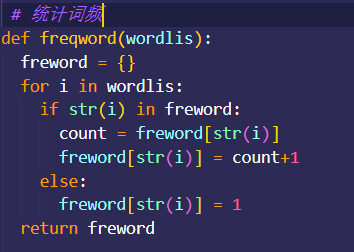
然后的程序逻辑为：读取数据、按照/切割成词表、获取停用词表、对比停用词表并去除相应的词、统计字典并返回相应词频、建立相应的语料库、查询文档数量、计算TF-IDF值、返回处理结果写入相应文件共九步。

2.py 为计算TF-IDF的python程序

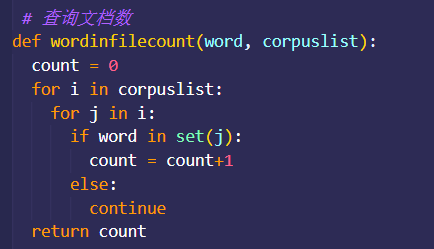
其中对比停用词表并去除相应的词：



统计词频：



查询文档数量：



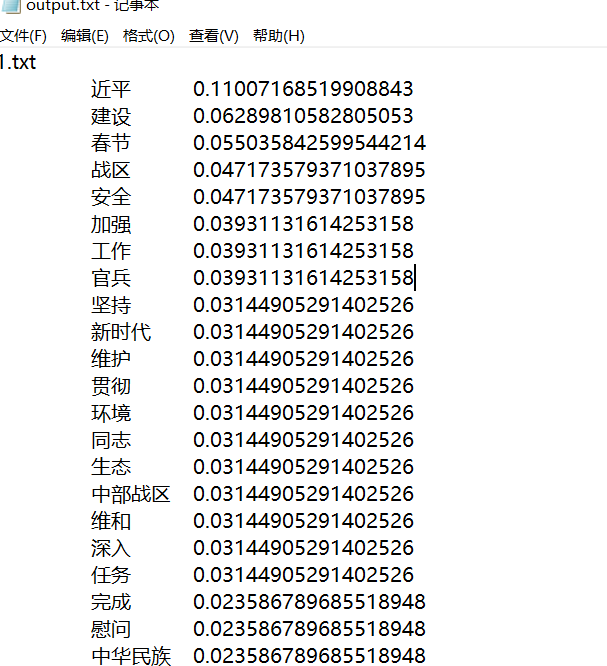
计算TF-IDF值：



main



最后所得关键词输出到output.txt文件中。



最终所得各文档关键词如下：

1.txt

近平 0.11007168519908843

建设 0.06289810582805053

春节 0.055035842599544214

战区 0.047173579371037895

2.txt

总书记 0.10127413415157764

总统 0.08680640070135226

近平 0.07957253397623958

生态 0.04340320035067613

3.txt

总书记 0.15378013335999435

近平 0.07097544616615124

年馍 0.047316964110767495

政府 0.047316964110767495

4.txt

总书记 0.11446530288560666

老蔡 0.06359183493644815

收益 0.03815510096186889

脱贫 0.03815510096186889

5.txt

总书记 0.12513891633796226

生活 0.06256945816898113

老蔡 0.050055566535184906

农村 0.050055566535184906

6.txt

克强 0.10886210624085668

发展 0.07257473749390446

国家 0.0635028953071664

企业 0.0635028953071664

7.txt

生态 0.1548648215511149

建设 0.10773204977468863

文明 0.10099879666377058

工作 0.06059927799826235

8.txt

生态 0.13368848352288412

发展 0.10397993162890987

环境 0.0965527936554163

资源 0.08169851770842919

9.txt

发展 0.1208639844133735

生态 0.05640319272624096

建设 0.05640319272624096

推动 0.048345593765349394

10.txt

发展 0.1159574018059032

生态 0.07247337612868951

绿色 0.06522603851582055

自然 0.06522603851582055

11.txt

生态 0.12478782518991988

攻坚战 0.08579162981806993

发展 0.07799239074369993

建设 0.06239391259495994

12.txt

发展 0.1263065411151522

生态 0.11841238229545517

保护 0.08683574701666712

黄河 0.0631532705575761

13.txt

发展 0.1452742350964674

生态 0.08410613610848112

治理 0.07646012373498283

坚持 0.061168098987986266

14.txt

发展 0.08648489551356948

克强 0.06289810582805053

推动 0.047173579371037895

变化 0.047173579371037895

15.txt

工作 0.06010533278935331

运行 0.052592166190684145

北京 0.04507899959201497

近平 0.04507899959201497

16.txt

乡村 0.05548419976939794

群众 0.05548419976939794

总书记 0.05548419976939794

脱贫 0.05548419976939794

17.txt

能源 0.07752630075917583

推动 0.06977367068325825

安全 0.062021040607340656

山西 0.054268410531423075

18.txt

表示 0.08324749300771393

冬奥会 0.06937291083976162

防控 0.06243561975578546

中国 0.05549832867180929

19.txt

雪橇 0.11165058232284585

北京 0.07975041594488988

中方 0.07177537435040089

蒙方 0.06380033275591192

20.txt

北京 0.14572895389401025

冬奥会 0.1285843710829502

运动 0.09429520546083016

中国 0.06857833124424012

21.txt

计量 0.08309936579595643

发展 0.08309936579595643

推动 0.04847463004764126

工作 0.04847463004764126

22.txt

计量 0.13232189013576132

发展 0.13232189013576132

建设 0.10118732775087629

航天 0.10118732775087629

23.txt

探测 0.08154290148422265

实施 0.05930392835216193

月球 0.05189093730814169

发射 0.044477946264121444

24.txt

总书记 0.1531503358052793

近平 0.05405305969598092

重建 0.045044216413317434

…… 0.03603537313065395

25.txt

总书记 0.14496751826773108

近平 0.07805943291339366

脱贫 0.05575673779528118

乡亲 0.044605390236224944

26.txt

总书记 0.13016121398363298

老蔡 0.09111284978854309

帮扶 0.0520644855934532

收益 0.0390483641950899

27.txt

总书记 0.07054659150653704

发展 0.058788826255447534

中国 0.04703106100435803

生活 0.04703106100435803

28.txt

航天 0.14382814145191447

发展 0.14382814145191447

空间 0.06768383127148915

建设 0.059223352362553015

29.txt

发射 0.12749066493810673

卫星 0.11407059494462181

系统 0.09394048995439443

成功 0.07381038496416706

30.txt

发射 0.0835453373476342

探测 0.06962111445636184

通信 0.05569689156508947

卫星 0.048734780119453286

31.txt

应用 0.09035735352163975

空间 0.08390325684152262

服务 0.0774491601614055

建设 0.07099506348128838

32.txt

卫星 0.08077362882871111

空间 0.08077362882871111

发展 0.06608751449621818

科学 0.051401400163725255

33.txt

航天 0.15262040384747555

建设 0.06867918173136399

发展 0.06867918173136399

商业航天 0.06104816153899022

34.txt

—— 0.1292737511610173

合作 0.1292737511610173

国际 0.12246986952096378

空间 0.10205822460080315

35.txt

—— 0.14321898697047106

卫星 0.09963059963163204

空间 0.09963059963163204

国家 0.08717677467767804

36.txt

—— 0.14908182593568933

卫星 0.08631053080487276

合作 0.07846411891352069

推动 0.07061770702216863

37.txt

浪费 0.11498559971690488

餐饮 0.05307027679241763

节约 0.04422523066034803

簕竹镇 0.03538018452827842

38.txt

航天 0.062483353410033

餐饮 0.05355716006574257

浪费 0.05355716006574257

平台 0.04463096672145214

39.txt

发展 0.1426736565498804

航天 0.11096839953879588

空间 0.07926314252771134

推动 0.06341051402216906

40.txt

发射 0.12098923206561533

卫星 0.10082436005467943

成功 0.10082436005467943

能力 0.07393786404009826

41.txt

发射 0.08828325818020326

探测 0.06621244363515243

技术 0.06621244363515243

实施 0.058855505453468834

42.txt

卫星 0.08753229044193452

服务 0.08753229044193452

应用 0.08079903733101647

发展 0.08079903733101647

43.txt

科学 0.0821059134622495

卫星 0.057474139423574655

航天 0.057474139423574655

发展 0.0492635480773497

44.txt

航天 0.13023808084228475

加强 0.06128850863166341

卫星 0.05362744505270548

商业航天 0.045966381473747556

45.txt

—— 0.15212326898152284

合作 0.12043092127703893

卫星 0.10775398219524535

国际 0.0824001040316582

46.txt

—— 0.13547475721271168

卫星 0.11700092668370554

合作 0.11700092668370554

国际 0.11084298317403683

47.txt

—— 0.0659630559001801

运动员 0.05130459903347342

光伏 0.04397537060012007

冬奥村 0.04397537060012007

48.txt

冬奥村 0.10810611939196184

北京 0.09979026405411863

运动员 0.06652684270274575

北京冬奥村 0.04157927668921609

49.txt

运动员 0.10211773376966658

传递 0.0907713189063703

冬奥村 0.07942490404307401

北京 0.07942490404307401