C#实验报告

—————基于脑癌的趋势分析

学院: 软件学院

年级: 2018

班级: 10

1. **成员及分工**

55181022崔益昊：占比35%，负责：python编写爬虫，爬取文章名字、时间、作者关键词和摘要，编写TF-IDF算法，提取每个文章的关键词并写入文件中，撰写实验报告。

55181015刘泓见：占比25%，负责：读取文章关键词并统计词频，并根据年，季度，月统计不同的关键词词频。

55181025冀耕辉：占比25%，负责： 读取爬虫文件内容，研究文献关注度。

55181027董昊：占比15%，负责：总结数据，运用html，css，JavaScript相关知识构建网页，通过在echarts官网下载相关插件，修改实例的option函数，在对应的六个具备高宽的DOM容器中引入chart表格构建网页细节展示数据。

二、摘要

通过关键字检索，从网站中读取脑癌11年内的数据文件，以及合计460篇文献，读入txt文档中，读取文件，分析时间和关键词，通过趋势得出结论

网站信息：

http://apps.webofknowledge.com/full\_record.do?product=UA&search\_mode=AdvancedSearch&qid=472&SID=7Ct8pADGmJulJgDoC9Z&page=0&doc=

三、导言

在对该癌症进行数据分析之前，我们了解了基本病情，查阅相关信息，用python编写爬虫爬取文章标题、时间、摘要和关键词等数据，然后读取爬虫文件内容，得出学术关注度和文件关键词，基于文本的关键词提取技术生成月，季度，年的词云。

四、方法

TF-IDF

def tf\_idf(wordlis, filelist, corpuslist): # 计算TF-IDF,并返回字典

outdic = {}

tf = 0

idf = 0

dic = freqword(wordlis)

# outlis = []

for i in set(wordlis):

tf = dic[str(i)] / len(wordlis) # 计算TF：某个词在文章中出现的次数/文章总词数

# 计算IDF：log(语料库的文档总数/(包含该词的文档数+1))

idf = math.log(len(filelist) / (wordinfilecount(str(i), corpuslist) + 1))

tfidf = tf \* idf # 计算TF-IDF

outdic[str(i)] = tfidf

orderdic = sorted(outdic.items(), key=operator.itemgetter(1),reverse=True) # 给字典排序

return orderdic

分词函数

def participles(text): # 分词函数

pattern = r"""(?x) # set flag to allow verbose regexps

(?:[A-Z]\.)+ # abbreviations, e.g. U.S.A.

|\d+(?:\.\d+)?%? # numbers, incl. currency and percentages

|\w+(?:[-']\w+)\* # words w/ optional internal hyphens/apostrophe

|\.\.\. # ellipsis

|(?:[.,;"'?():-\_`]) # special characters with meanings

"""

t = nltk.regexp\_tokenize(text, pattern)

length = len(t)

for i in range(length):

t[i] = t[i].lower()

return t

去除文件停用词

def getridofsw(lis, swlist): # 去除文章中的停用词

afterswlis = []

for i in lis:

if str(i) in swlist:

continue

else:

afterswlis.append(str(i).lower())

return afterswlis

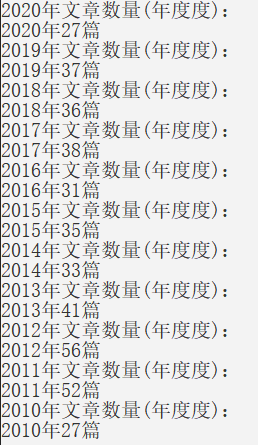
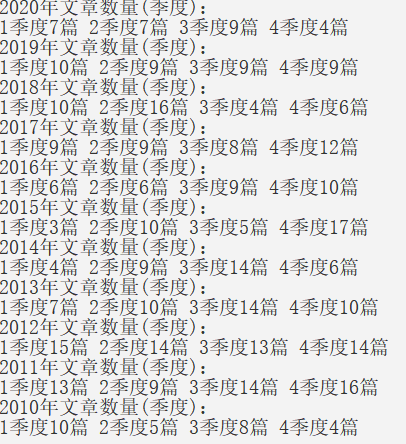
利用哈夫曼树统计词频

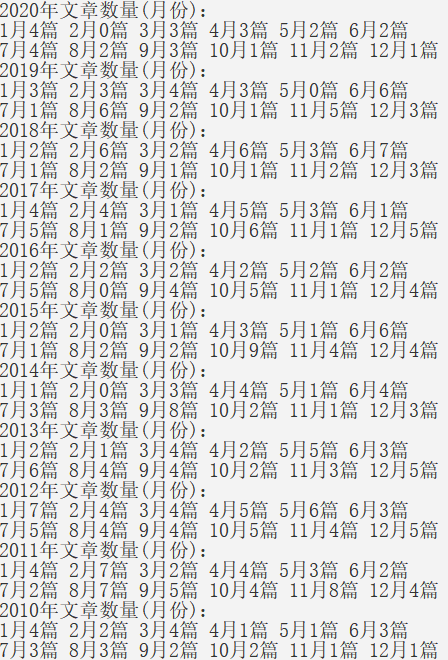
private Hashtable hashcreate(ArrayList strword)  
{  
  
Hashtable ht = new Hashtable();  
foreach (string i in strword)//遍历wtrword  
if (!ht.Contains(i))  
ht.Add(i, 1);  
else  
{  
ht[i] = (int)ht[i] + 1;  
}  
return ht;  
}

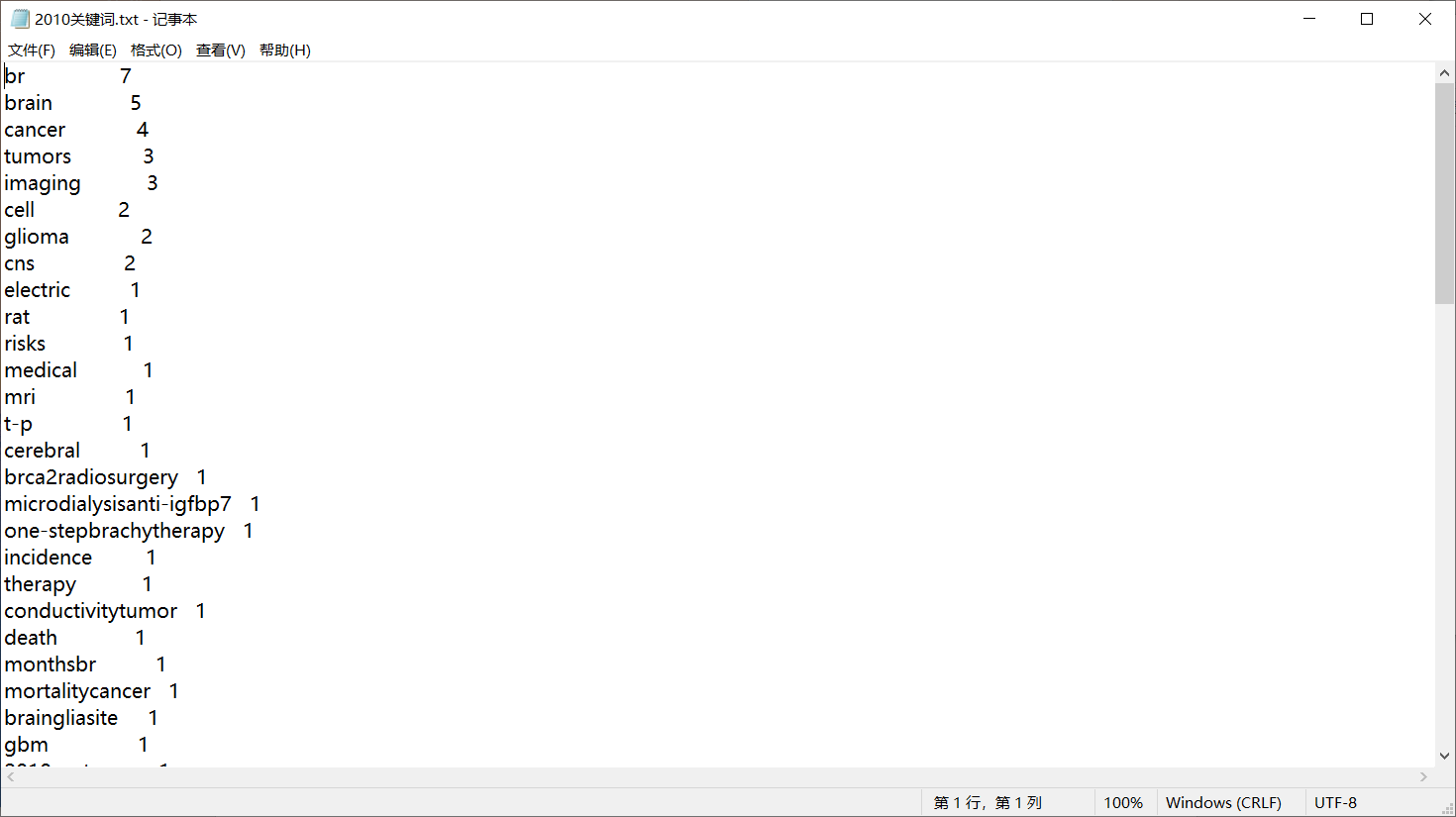
分隔单词

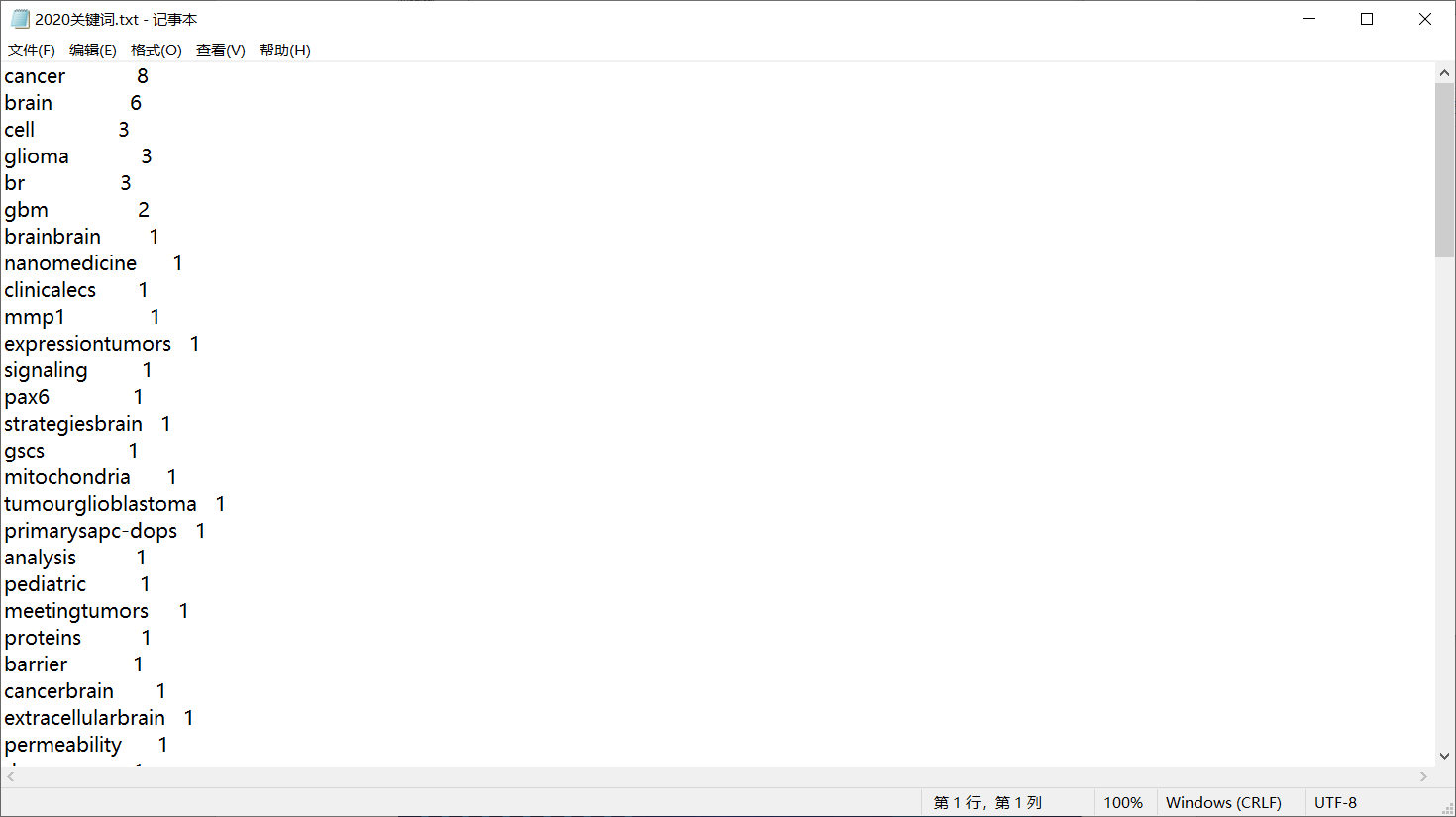
private ArrayList separateword(string strtext\_low)  
{  
  
ArrayList sepword = new ArrayList();//定义一个动态数组用于传进已转换小写的数组  
string strtext = strtext\_low;  
  
  
String[] word = strtext.Split(&apos;[&apos;, &apos;]&apos;, &apos;,&apos;, &apos;\&apos;&apos;,&apos;<&apos;,&apos;>&apos;,&apos;=&apos;,&apos;"&apos;);//使用Split（）方法分割单词  
foreach (string i in word)//遍历循环将其传入动态数组sepword  
sepword.Add(i);  
return sepword;  
}

五、数据分析和结果







图用echart等技术实现，搭建网站，将多组图片构建如网页中显示，使结果可以通过网页浏览。



六、结论

2011和2012研究脑癌文献数量是最多的，研究重点主要是Glioblastoma,Intracranial tumors,cells, intracranial,文章数量增长趋势比较平稳，每个季度的文章都在100左右，月份上下有浮动，但是浮动不大。

七、参考资料

1. **EGFRvIII expression and isocitrate dehydrogenase mutations in patients with glioma**

**作者:**[Taher, MM](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=32418918) (Taher, Mohiuddin M.)**[**[**1**](javascript:sup_focus('addressWOS:000595649300046-1'))**,**[**2**](javascript:sup_focus('addressWOS:000595649300046-2'))**]**; [Dairi, G](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=42252708" \o "查找此作者的更多记录) (Dairi, Ghida)**[**[**3**](javascript:sup_focus('addressWOS:000595649300046-3'))**,**[**4**](javascript:sup_focus('addressWOS:000595649300046-4'))**]**; [Butt, EM](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=42254896) (Butt, Ejaz Muhammad)**[**[**5**](javascript:sup_focus('addressWOS:000595649300046-5'))**,**[**6**](javascript:sup_focus('addressWOS:000595649300046-6'))**]**; [Al-Quthami, K](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=33733892) (Al-Quthami, Khalid)**[**[**5**](javascript:sup_focus('addressWOS:000595649300046-5'))**,**[**6**](javascript:sup_focus('addressWOS:000595649300046-6'))**]**; [Al-Khalidi, H](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=1350915) (Al-Khalidi, Hisham)**[**[**7**](javascript:sup_focus('addressWOS:000595649300046-7'))**]**; [Jastania, RA](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=5091909" \o "查找此作者的更多记录) (Jastania, Raid A.)**[**[**8**](javascript:sup_focus('addressWOS:000595649300046-8'))**,**[**9**](javascript:sup_focus('addressWOS:000595649300046-9'))**]**; [Nageeti, TH](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=4906740" \o "查找此作者的更多记录) (Nageeti, Tahani H.)**[**[**10**](javascript:sup_focus('addressWOS:000595649300046-10'))**]**; [Bogari, NM](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=5156082" \o "查找此作者的更多记录) (Bogari, Neda M.)**[**[**1**](javascript:sup_focus('addressWOS:000595649300046-1'))**]**; [Athar, M](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=23718) (Athar, Mohammad)**[**[**1**](javascript:sup_focus('addressWOS:000595649300046-1'))**,**[**2**](javascript:sup_focus('addressWOS:000595649300046-2'))**]**; [Al-Allaf, FA](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=36205597) (Al-Allaf, Faisal A.)**[**[**1**](javascript:sup_focus('addressWOS:000595649300046-1'))**,**[**2**](javascript:sup_focus('addressWOS:000595649300046-2'))**]**; [Valerie, K](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=170029) (Valerie, Kristoffer)**[**[**11**](javascript:sup_focus('addressWOS:000595649300046-11'))**,**[**12**](javascript:sup_focus('addressWOS:000595649300046-12'))**]**

2. **owards the targeted assessment of relevant problems: Optimization of the distress Thermometer for adult neuro-oncological patients**

**作者:**[Goebel, S](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=2703297) (Goebel, Simone)**[**[**1**](javascript:sup_focus('addressWOS:000577737000001-1'))**]**; [Knuth, C](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=41818360) (Knuth, Carina)**[**[**1**](javascript:sup_focus('addressWOS:000577737000001-1'))**]**; [Damm, M](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=8490655" \o "查找此作者的更多记录) (Damm, Matthias)**[**[**2**](javascript:sup_focus('addressWOS:000577737000001-2'))**]**; [Linden, D](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=34639783) (Linden, Daniel)**[**[**2**](javascript:sup_focus('addressWOS:000577737000001-2'))**]**; [Coburger, J](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=41814703" \o "查找此作者的更多记录) (Coburger, Jan)**[**[**3**](javascript:sup_focus('addressWOS:000577737000001-3'))**]**; [Ringel, F](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=36225032" \o "查找此作者的更多记录) (Ringel, Florian)**[**[**4**](javascript:sup_focus('addressWOS:000577737000001-4'))**]**; [Tabatabai, G](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=251146" \o "查找此作者的更多记录) (Tabatabai, Ghazaleh)**[**[**2**](javascript:sup_focus('addressWOS:000577737000001-2'))**]**; [Mehdorn, M](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=30352756) (Mehdorn, Maximilian)**[**[**5**](javascript:sup_focus('addressWOS:000577737000001-5'))**]**; [Renovanz, M](http://apps.webofknowledge.com/OutboundService.do?SID=8BGZMT3brJ1tk7YT1js&mode=rrcAuthorRecordService&action=go&product=WOS&lang=zh_CN&daisIds=41824746" \o "查找此作者的更多记录) (Renovanz, Mirjam)**[**[**2**](javascript:sup_focus('addressWOS:000577737000001-2'))**,**[**6**](javascript:sup_focus('addressWOS:000577737000001-6'))**]**

3.CSDN https://blog.csdn.net/whyoceansea/article/details/79808854?ops\_request\_misc=&request\_id=&biz\_id=102&utm\_term=python+%25E6%258F%2590%25E5%258F%2596%25E8%258B%25B1%25E6%2596%2587%25E5%2585%25B3%25E9%2594%25AE%25E8%25AF%258D&utm\_medium=distribute.pc\_search\_result.none-task-blog-2~blog~sobaiduweb~default-7-79808854.pc\_v1\_rank\_blog\_v1