基于SCI的胃癌趋势分析

成员：侯钧哲 100%

**摘要：**通过对SCI近10年的有关胃癌文献数据的分析，统计各年度、季度文献篇数来确定学术关注度，并根据文献的摘要分析出胃癌研究现状并生成对应的年度、季度词云，通过打开指定网页进行查看。根据这些近10年的数据直观显示胃癌趋势以及胃癌的研究发展。

**关键词：**胃癌，趋势分析。

**介绍：**通过C#程序编写爬虫获取SCI中有关胃癌的文献数据，并对每一个文献的摘要进行分析。根据年度、季度、月度分别生成词云并在浏览器中显示，同时统计各年度的文献数目，根据词云的密集度及各年的文献数目分析学术关注度趋势，得到最后的结论。

**方法：**利用C#的WinForm开发框架，用WebBrowser打开动态网页以便于获取网页内数据。根据SCI网站网页的特点，每次翻页只需要将url地址中page参数设置为下一页。首先分析检索结果，获取网页中文献地址以及文献发布的时间，并从其中的时间中提取出年份、月份（如果有的话），对某些只提供出年份时间的文献，将其划分为第0月，不参与月度、季度词云生成，但参与年度词云生成，季度的划分按三个月一季，从一月份开始。

由于文献中的数据格式不统一，有的月份用数字表示，有的按英文单词前三个的大写或首字母大写的形式，有的直接用季度如Aut、Fal、Win、FAL等表示，有的甚至没有标明月份。需要对这些不统一的格式进行统一化处理，利用C#中的枚举将这些标识统一以数字表示，并生成一个以月份为键的字典来统计该年中每月的词频。

查询的结果中默认是按时间排序的（偶尔会有偏差，但基于年份的偏差较小），可以根据是否年份发生变化来确定是否进入下一年的统计，如果将进入下一年的统计，则会对前一年的数据进行处理，生成对应的年度、季度、月度词云等，之后清空前一年的数据以防止内存占用过大，同时也使每年的统计更清晰。

由于需要查询近10年的数据，数据量过大而且爬虫爬取一定时间后就会被网站封锁IP，为了确保未完成的查询过程的时间不被浪费，每当对10页的数据分析之后就会在break文件夹下创建一个page.txt来存储当前页数，words\_key存储已经查询到的月份，words\_value对应存储每月的词及词频，years\_key存储已经查询的年份，years\_value对应存储该年已分析的文献数。每次程序开始时会先查询page.txt文件是否存在，若存在则读取page以及words、years的数据接续上一次该点的查询。每一次的对这些文件的更改都会覆盖之前的数据。当未能访问到有关数据时（例如暂时的IP被封）会停止继续查询，发出信息提醒并结束运行。

对文献进行分析的部分只是对摘要的简单分析，去掉前导换行空格以及句中标点等，之后提取摘要中的文字，并对其中存在的某些设置的stop停顿词进行筛除，并将结果写入该月的字典中统一进行统计。

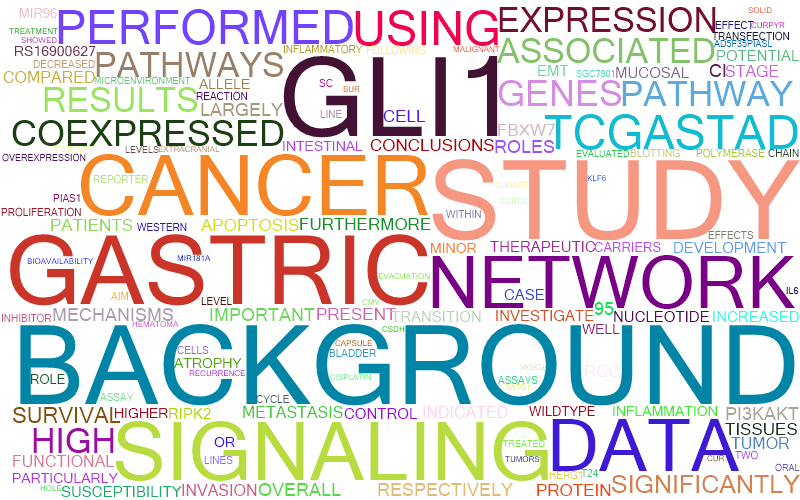
在词云生成部分，对数据会进一步的按照不同的方式进行筛选。对于月度词云，会选出本月里词频至少为3的词，对于季度和年度词云，依据季度词频筛选=本年的词数/1000、年度词频筛选=本年的次数/500进行计算，生成对应以a为开头的年度词云，b为开头的季度词云，c为开头的月度词云并保存在image文件夹下，便于最终在网页中显示。每完成一个词云就会清除该年中的词频等数据。

在最终的显示部分，当程序完成10年内的查询时，将打开Key.html网页显示有关的数据。在该网页里，将以折线图形式显示年度的文献数，并在具体某年中显示季度和月度词云。

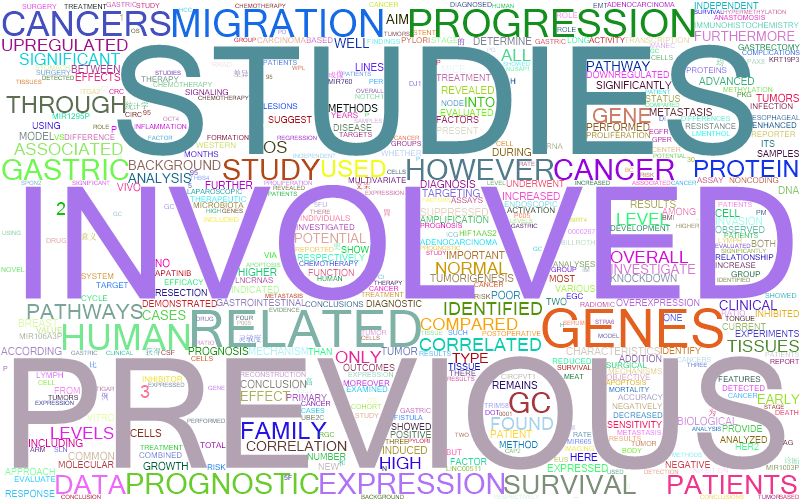
**数据分析及结果：**

以下是其中的一些示例结果：

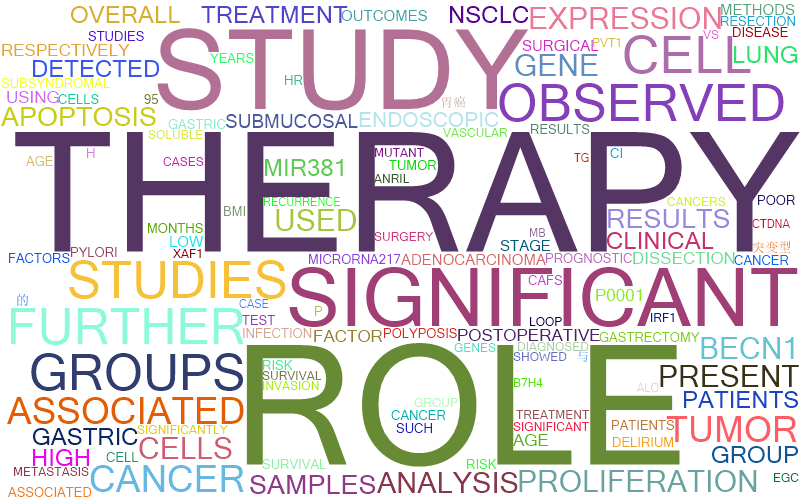
**2018年3月份的词云**

****

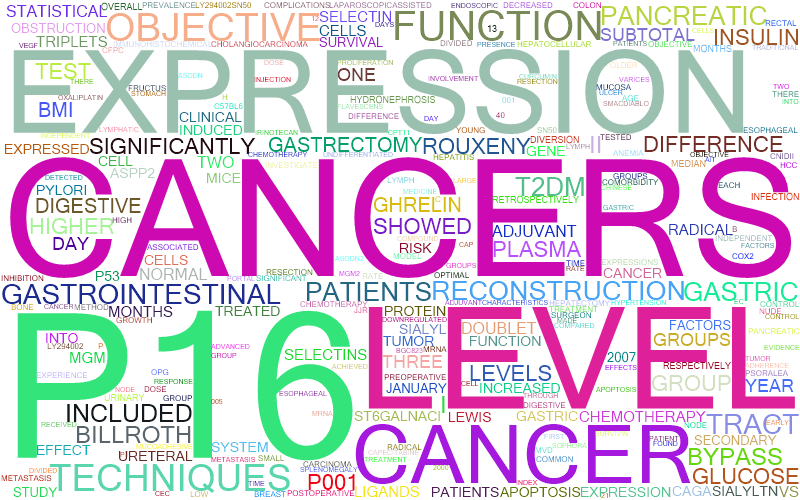
**2019年的年度词云：**

****

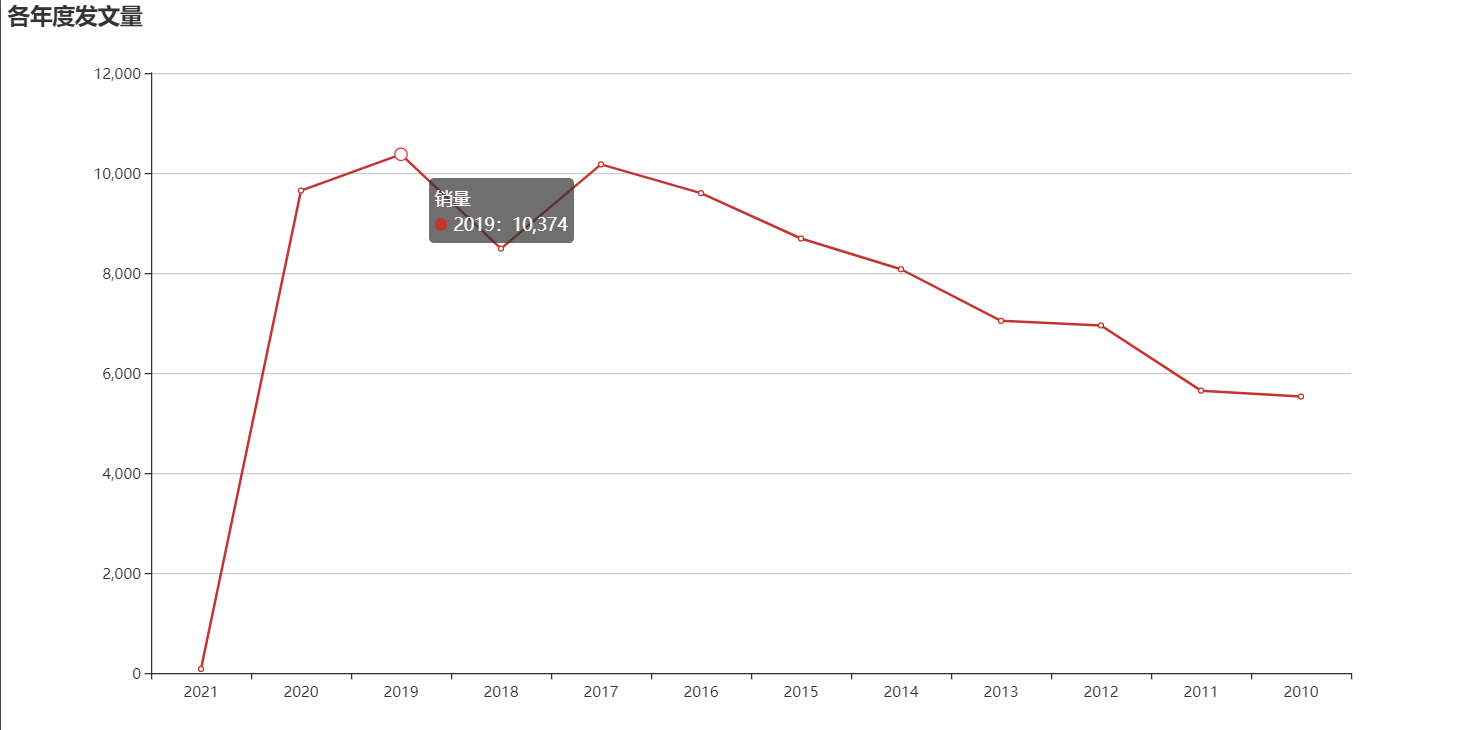
**2018年第三季度的词云：**

****

**2011年1季度词云：**

****

**Echart实现年度图表：**

****

**结论：**

从文献数目上看，从2010年到2021年，每年新发文献数目呈上升趋势 ，近年来维持在9000篇左右，可见人们对胃癌的研究越来越重视，对胃癌的研究也逐渐的取得进展。

**参考文献：**

[1]周宁，石雯茜、朱昭昭，基于粗糙数据推理的TextRank关键词提取算法

[2]肖辛格．基于基本层次范畴改进TextRank算法的中文关键词抽取[D]．武汉：华中师范大学硕士学位论文，2017．

[3]刘竹辰，陈浩，于艳华，等．词位置分布加权TextRank的关键词提取[J]．数据分析与知识发现，2018,2(9):74-79．

[4]Turney P D.Learning algorithms for keyphrase extraction[J].Information Retrieval,2000,2(4)