**2018级软件学院c#研究报告**

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 学号 | 班级 | 占比 |
| 张宇浩 | 65180308 | 5 | 100% |

**摘要**：此报告的主题是基于文献的癌症趋势分析，通过c#语言，对特定网站进行爬虫，选择了三种癌症共7500条论文数据进行统计，具体统计方法为：按照年、季度、月的时间对论文的数量进行调查。最终得出有关癌症的文献研究趋势。

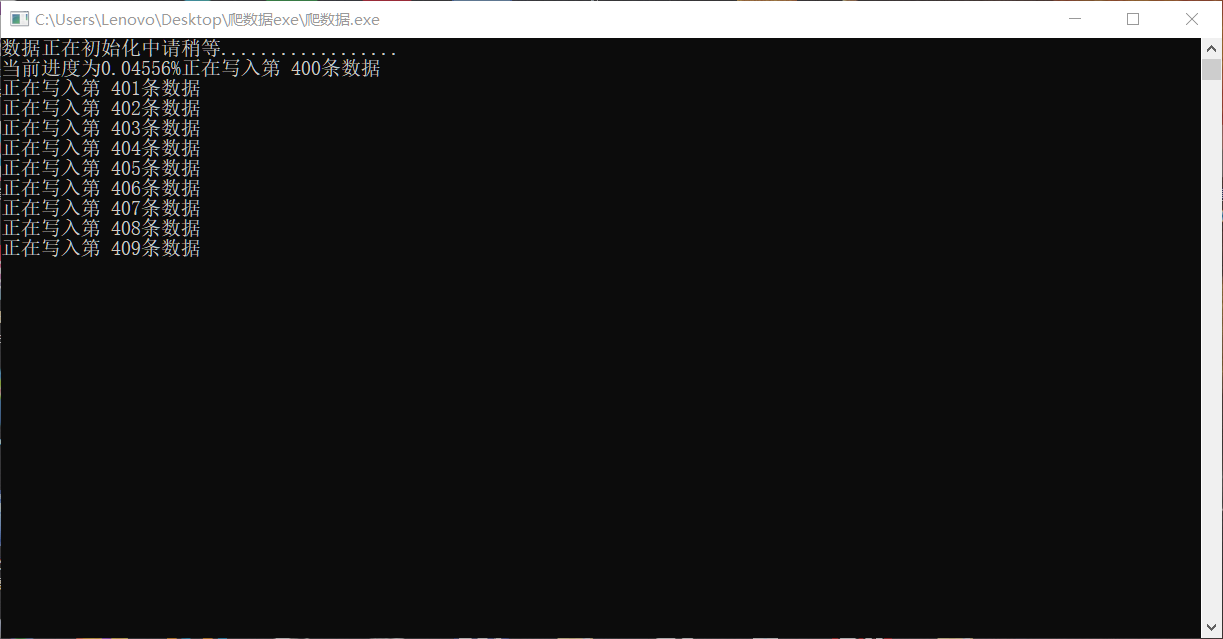
1. **Introduction**

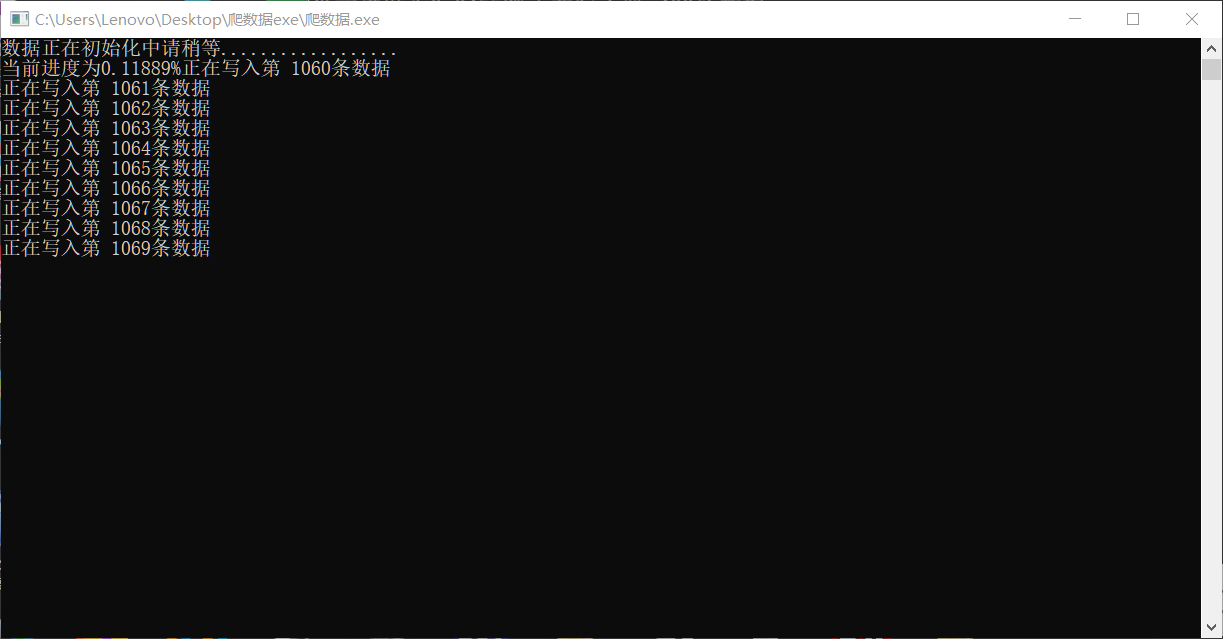
本研究基于c#对有关癌症的文献进行趋势分析，分别选择了肺癌、胃癌、乳腺癌三种癌症共7500篇论文进行年、季度、月的统计，并得到了相关的结论。

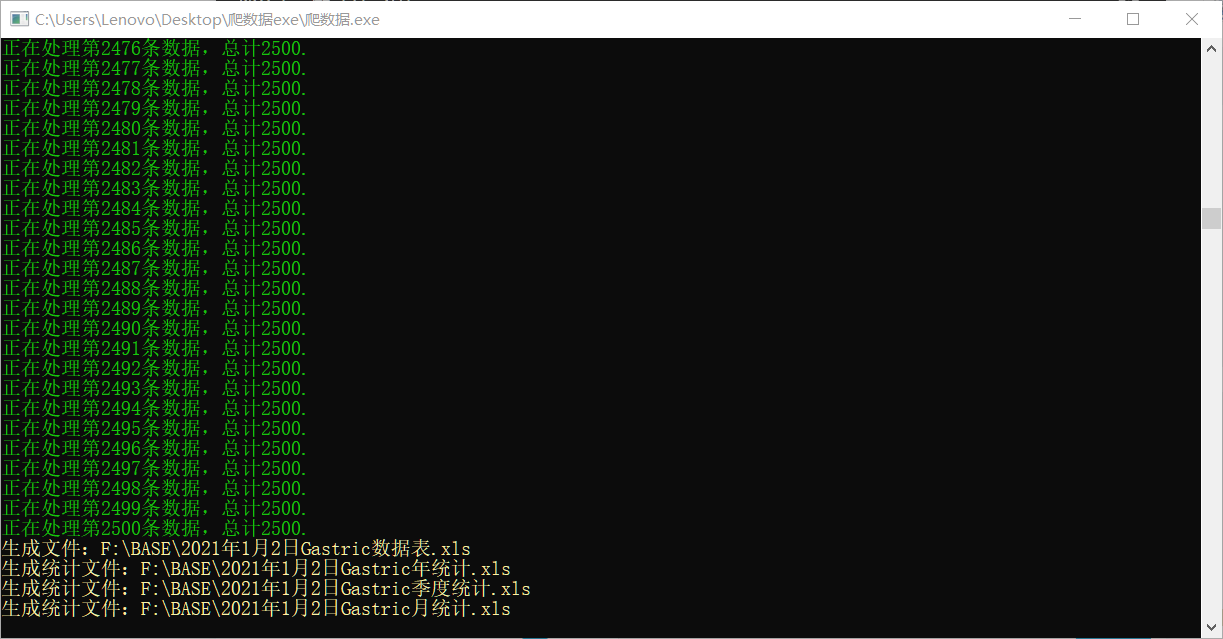
1. **Review of Literature**

在项目中，我们首先通过编写c#语言对论文网站进行了爬虫，分别在肺癌、胃癌、乳腺癌中爬取了2500篇论文，每篇论文为近10年中随机抽取得到，并将年、季度、月的数据深入分析形成了词云。具体地，我们以癌症名字为关键词检索，并考虑了同义词的情况，得到了相关论文。然后，利用c#进行爬虫，对这些论文出版的年、季度、月进行了统计，并且保证数据不少于10年。通过以上的分析，我们得到了这些论文的学术关注度，研究重点，研究趋势等方面的信息。

具体研究过程如下图：







得到的词云如下：（季度词云中的1、2、3、4分别代表春夏秋冬）

**肺癌：**



年词云



季度词云

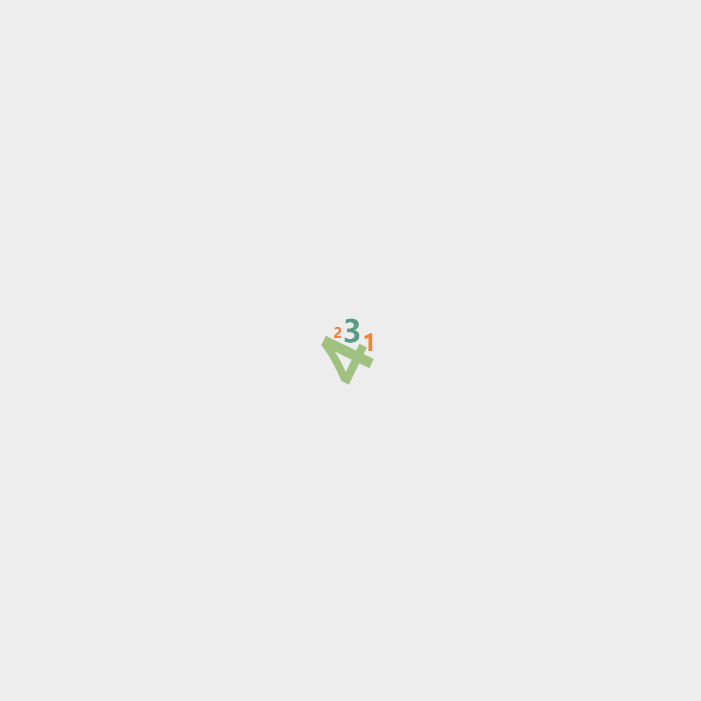


月词云

**胃癌：**



年词云



季度词云



月词云

**乳腺癌：**



年词云



季度词云



月词云

1. **Methods**

首先通过Console.SetCursorPositon()设置光标位置，然后通过Properties.Settings.Default.URL设置关键字，分别为”Lung”,”Breast”,”Gastric”,进而通过httpGet获取请求，如果搜索到的字符串为空就跳出当前循环，进入下一个循环。对于月份的处理是，利用UI.TIME.Contains进行月份的匹配，若成功则通过split函数将月份单独取出。最后将得到的数据按照癌症类别进行分类，根据年、季度、月各分为三个部分，存入excel表格中。HttpGet的实现过程如下：

public static string httpGet(string URL)

{

HttpWebRequest req = (HttpWebRequest)WebRequest.Create(URL);

req.Proxy = null;

WebResponse resp = req.GetResponse();

StreamReader sr = new StreamReader(resp.GetResponseStream(), Encoding.Default);

try

{

string sReturn = sr.ReadToEnd();

resp.Close(); sr.Close();

return sReturn;

}

catch

{

return "";

}

}

1. **Data Analysis and Results**

所得到的数据如下：

1. **肺癌：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **名称** | **年份** | **季度** | **月份** | **数量** |
| 1 | Lung Cancer | 2020 |  |  | 777 |
| 2 | Lung Cancer | 2019 |  |  | 298 |
| 3 | Lung Cancer | 2018 |  |  | 218 |
| 4 | Lung Cancer | 2017 |  |  | 188 |
| 5 | Lung Cancer | 2016 |  |  | 211 |
| 6 | Lung Cancer | 2015 |  |  | 146 |
| 7 | Lung Cancer | 2014 |  |  | 132 |
| 8 | Lung Cancer | 2013 |  |  | 198 |
| 9 | Lung Cancer | 2012 |  |  | 111 |
| 10 | Lung Cancer | 2011 |  |  | 61 |
| 11 | Lung Cancer | 2010 |  |  | 82 |
| 12 | Lung Cancer | 2009 |  |  | 44 |
| 13 | Lung Cancer | 2008 |  |  | 15 |
| 14 | Lung Cancer | 2007 |  |  | 13 |
| 15 | Lung Cancer | 2006 |  |  | 3 |
| 16 | Lung Cancer | 2005 |  |  | 3 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Lung Cancer |  | 4 |  | 723 |
| 2 | Lung Cancer |  | 3 |  | 657 |
| 3 | Lung Cancer |  | 2 |  | 527 |
| 4 | Lung Cancer |  | 1 |  | 593 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Lung Cancer |  |  | 12 | 283 |
| 2 | Lung Cancer |  |  | 11 | 233 |
| 3 | Lung Cancer |  |  | 10 | 207 |
| 4 | Lung Cancer |  |  | 9 | 196 |
| 5 | Lung Cancer |  |  | 8 | 198 |
| 6 | Lung Cancer |  |  | 7 | 263 |
| 7 | Lung Cancer |  |  | 6 | 185 |
| 8 | Lung Cancer |  |  | 5 | 190 |
| 9 | Lung Cancer |  |  | 4 | 152 |
| 10 | Lung Cancer |  |  | 3 | 199 |
| 11 | Lung Cancer |  |  | 2 | 208 |
| 12 | Lung Cancer |  |  | 1 | 186 |

1. **胃癌：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **名称** | **年份** | **季度** | **月份** | **数量** |
| 1 | Breast Cancer | 2020 |  |  | 844 |
| 2 | Breast Cancer | 2019 |  |  | 334 |
| 3 | Breast Cancer | 2018 |  |  | 188 |
| 4 | Breast Cancer | 2017 |  |  | 199 |
| 5 | Breast Cancer | 2016 |  |  | 207 |
| 6 | Breast Cancer | 2015 |  |  | 151 |
| 7 | Breast Cancer | 2014 |  |  | 115 |
| 8 | Breast Cancer | 2013 |  |  | 141 |
| 9 | Breast Cancer | 2012 |  |  | 108 |
| 10 | Breast Cancer | 2011 |  |  | 53 |
| 11 | Breast Cancer | 2010 |  |  | 44 |
| 12 | Breast Cancer | 2009 |  |  | 44 |
| 13 | Breast Cancer | 2008 |  |  | 22 |
| 14 | Breast Cancer | 2007 |  |  | 20 |
| 15 | Breast Cancer | 2006 |  |  | 6 |
| 16 | Breast Cancer | 2005 |  |  | 14 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Breast Cancer |  | 4 |  | 739 |
| 2 | Breast Cancer |  | 3 |  | 610 |
| 3 | Breast Cancer |  | 2 |  | 556 |
| 4 | Breast Cancer |  | 1 |  | 585 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Breast Cancer |  |  | 12 | 292 |
| 2 | Breast Cancer |  |  | 11 | 228 |
| 3 | Breast Cancer |  |  | 10 | 219 |
| 4 | Breast Cancer |  |  | 9 | 225 |
| 5 | Breast Cancer |  |  | 8 | 199 |
| 6 | Breast Cancer |  |  | 7 | 186 |
| 7 | Breast Cancer |  |  | 6 | 196 |
| 8 | Breast Cancer |  |  | 5 | 194 |
| 9 | Breast Cancer |  |  | 4 | 166 |
| 10 | Breast Cancer |  |  | 3 | 180 |
| 11 | Breast Cancer |  |  | 2 | 198 |
| 12 | Breast Cancer |  |  | 1 | 207 |

1. **乳腺癌：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **名称** | **年份** | **季度** | **月份** | **数量** |
| 1 | Gastric Cancer | 2020 |  |  | 884 |
| 2 | Gastric Cancer | 2019 |  |  | 363 |
| 3 | Gastric Cancer | 2018 |  |  | 205 |
| 4 | Gastric Cancer | 2017 |  |  | 203 |
| 5 | Gastric Cancer | 2016 |  |  | 193 |
| 6 | Gastric Cancer | 2015 |  |  | 135 |
| 7 | Gastric Cancer | 2014 |  |  | 87 |
| 8 | Gastric Cancer | 2013 |  |  | 178 |
| 9 | Gastric Cancer | 2012 |  |  | 96 |
| 10 | Gastric Cancer | 2011 |  |  | 62 |
| 11 | Gastric Cancer | 2010 |  |  | 41 |
| 12 | Gastric Cancer | 2009 |  |  | 28 |
| 13 | Gastric Cancer | 2008 |  |  | 12 |
| 14 | Gastric Cancer | 2007 |  |  | 11 |
| 15 | Gastric Cancer | 2006 |  |  | 1 |
| 16 | Gastric Cancer | 2005 |  |  | 1 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Gastric Cancer |  | 4 |  | 697 |
| 2 | Gastric Cancer |  | 3 |  | 669 |
| 3 | Gastric Cancer |  | 2 |  | 580 |
| 4 | Gastric Cancer |  | 1 |  | 554 |
| 序号 | 名称 | 年份 | 季度 | 月份 | 数量 |
| 1 | Gastric Cancer |  |  | 12 | 264 |
| 2 | Gastric Cancer |  |  | 11 | 215 |
| 3 | Gastric Cancer |  |  | 10 | 218 |
| 4 | Gastric Cancer |  |  | 9 | 215 |
| 5 | Gastric Cancer |  |  | 8 | 210 |
| 6 | Gastric Cancer |  |  | 7 | 244 |
| 7 | Gastric Cancer |  |  | 6 | 179 |
| 8 | Gastric Cancer |  |  | 5 | 207 |
| 9 | Gastric Cancer |  |  | 4 | 194 |
| 10 | Gastric Cancer |  |  | 3 | 165 |
| 11 | Gastric Cancer |  |  | 2 | 191 |
| 12 | Gastric Cancer |  |  | 1 | 198 |

数据在统计年份的过程中，有超过十年的冗余数据，本着统计的准确性，下述分析时将自动把冗余数据抛弃掉。

数据表明：在年份上，有关肺癌、胃癌、乳腺癌三种癌症的论文发表数量总体呈现上升的趋势，且都在2020年达到最大值，这里考虑到网站收录的最新论文较多且位置分布于前列，所以2020年论文的数量仅具备参考意义。

在季度上，三种癌症的发表数量均在第四季度达到最高，肺癌和胃癌在第二季度达到最低，乳腺癌在第一季度达到最低，但总体来看，三种癌症的论文发表数量均随季度的增大而增加，但相差不大。由此可见，季度对于论文的发表有一定影响，但不起到主要作用。

在月份上，三种癌症论文的发表数量均在12月达到顶峰，其他月份数量基本持平，可见月份对于论文的发表主要影响是12月份，其他月份影响较小。

1. **Conclusion**

年、季度、月对肺癌、胃癌、乳腺癌的论文发表数量都有一定的影响，但是影响各不相同，年份的影响较大，季度和月份有一定影响，但是不占主要因素。可见，随着年份的增加，论文数量也呈上升趋势，在这个角度说明，医药领域的科学研究在不断发展和壮大。

1. **Reference**