自然语言处理技术报告

郝久武

2022年9月23日

1 第一部分： 爬虫工具Selenium

Selenium是支持web浏览器自动化的一系列工具和库的综合项目。它提供了扩展来模拟用户与浏览器的交互，用于扩展浏览器分配的分发服务器，以及用于实现W3C WebDriver的基础结构，该规范允许用户为所有主要Web浏览器编写可互换的代码。

Selenium的核心是WebDriver，这是一个编写指令集的接口，可以在许多浏览器中互换运行。在Python中，使用Selenium示例如下：

图形用户界面, 文本, 应用程序, 聊天或短信

描述已自动生成

2 第二部分： 语料网站

为对比规范语料和非规范语料的差异，中文语料分别在**光明日报**（<https://epaper.gmw.cn/>）和**笔趣阁**（<https://www.bbiquge.net/>）两个网站上爬取得到，英文语料通过**英文小说网**（<http://novel.tingroom.com/>）爬取得到。

光明日报是由中共中央主办，以知识分子为主要读者对象的思想文化大报，其创刊于1949年6月16日，至今报纸发行量达100多万份。光明日报的读者主要分布在政府机关、企事业单位、国办高校等；是知识分子互相交流的学术平台，具有一定的权威性和广泛性，是企事业单位和高校有效的展示平台。

文本

描述已自动生成

图 1：光明日报网站版面

笔趣阁是一家提供小说在线阅读服务的门户网站，成立于2012年。该网站主要功能有小说搜索、在线阅读、下载以及网络文学创作等。



图2：笔趣阁网站版面

英文小说网创办于2005年，主要提供不同类型英文小说的电子资源，包括经典英文小说、名人传记、宗教小说、科幻小说和儿童小说等各种类型英文小说的在线阅读和下载渠道。



图3：英文小说网版面

3 第三部分： 样本清洗

主要采用**正则表达式**来进行样本的清洗工作。

正则表达式是计算机科学的一个概念，主要使用单个字符串来描述、匹配一系列符合某个句法规则的字符串。

Unicode官网地址：<https://home.unicode.org/>

re模块python官网地址：<https://docs.python.org/3/library/re.html>

参考上述资料，汉字的unicode编码范围为4e00到9fa5，英文的unicode编码范围为0041-005A, 0061-007A，因此样本清洗方法如下：

文本

描述已自动生成

清洗后的样本规模如表1所示。

表1：语料规模大小

|  |  |  |
| --- | --- | --- |
|  | 原始数据规模 | 清洗后数据规模 |
| **光明日报** | 86.64M | 50.99M |
| **笔趣阁** | 117.87M | 99.69M |
| **英文小说网** | 1.57G | 1.34G |

4 第四部分： 结果分析

每次将文本规模增加2M，计算文本规模扩大后的熵，中文语料和英文语料的计算结果分别如图4、图5所示。

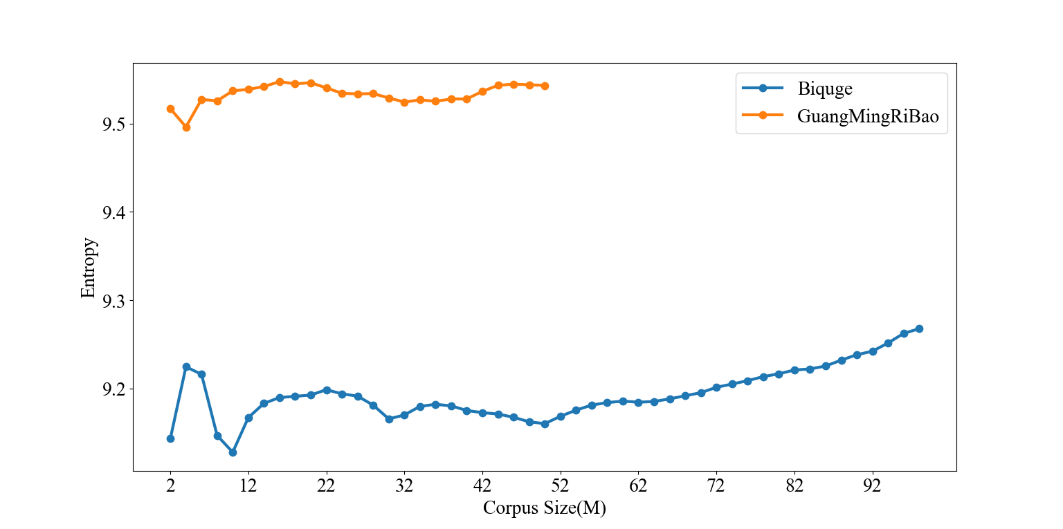


图4：中文语料中汉字的熵

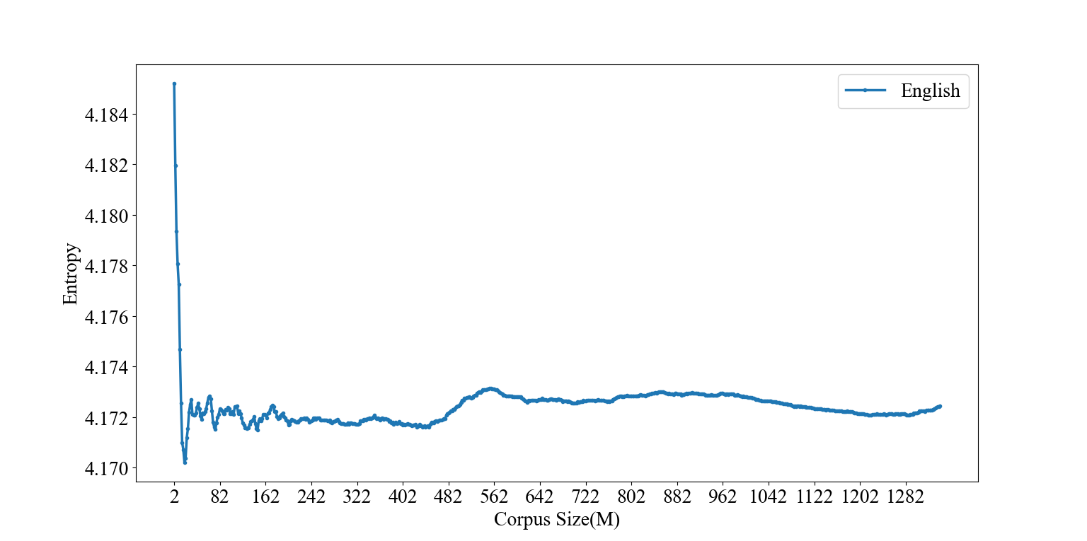


图5：英文语料中字母的熵

可以发现，随着文本规模的不断增大，文本的熵趋近于稳定。笔趣阁中文汉字的熵趋近于9.26，光明日报趋近于9.54，英语小说网的熵趋近于4.17，和PPT对比如表2所示。

表2：不同语料熵对比

|  |  |  |
| --- | --- | --- |
|  | 中文 | 英文 |
| **PPT数据** | 9.71 | 4.03 |
| **笔趣阁** | 9.26 | / |
| **光明日报** | 9.54 | / |
| **英语小说网** | / | 4.17 |

由于笔趣阁语料中经常出现主人公或生活的地名，因此这些特定词出现频率增多，整体的熵变小；而光明日报头版文章中较为规范，内容并无特定取向，因此汉字的熵较大。

表3：中文语料对比

|  |  |  |  |
| --- | --- | --- | --- |
|  | 出现汉字总字数 | 出现频率top5的汉字 | 出现频率top10的汉字 |
| **笔趣阁语料** | 5172 | **的**,是,一,了,不 | 的,是,一,了,不,  这,有,在,他,人 |
| **光明日报语料** | 6013 | **的**,国,中,一,人 | 的,国,中,一,人,  和,大,在,是,发 |

通过表3可以得知，常用汉字总字数为6000字左右，“的”是使用次数最多的汉字。