

**自 然 语 言 处 理**

**技术报告**

**班级：**2023春季自然语言处理班

**组号：--**

**姓名：**唐嘉良

**学号：**2020K8009907032

**报告主题：**计算中英文信息源的熵

**2023年4月2日**

**一．报告摘要**

**本次任务中，我采用python语言，利用BeautifulSoup等库进行爬虫、数据清洗和熵计算。**

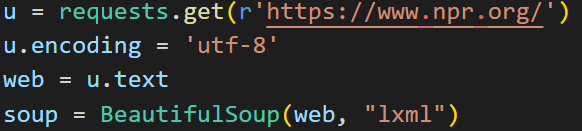
**针对中文熵计算，选取了https://news.163.com/（网易新闻）, [https://baike.baidu.com/（百度百科）,](https://baike.baidu.com/,) <http://www.xbiquge.la/10/10489/（笔趣阁）三个主网站；>**

**针对英文熵计算，选取了https://www.npr.org/, [https://english.cctv.com/,](https://baike.baidu.com/,) https://www.audible.com/三个主网站。**

**随后利用双层DFS搜索算法来检索次级网站并爬取数据，采用正则表达式匹配方法进行数据清洗。最后通过汉字统计计算中文熵、通过英文字母和空格统计计算英文熵。**

**二．网络爬虫与浅层DFS爬虫算法**

**采用requests, BeautifulSoup等库进行爬虫，用request.get函数获取网页内容，指定编码并利用BeautifulSoup解析格式。具体代码如下：**



**不同的网页有不同的特征，网址主页往往包括海量的次级链接。基于这一观察，采用如下2层DFS算法搜索主网页存放的次级网页网址信息。**

**首先搜索主网页所有子网页信息：**



**将子网页作为主网页，依次再度搜寻其子网页信息：**



**搜寻过程中以max\_num控制每一层搜寻广度上界。调整max\_num可以以更高的时间开销爬虫更多文本。此外，搜寻过程中需要进行网址去重，确保网页信息不会重复爬取。**

**某些特殊网页的特征明显。例如，针对小说网站目录页类型的主页，其次级网页一般不会有太多信息质量较高的次级网页，只需进行一层搜索，确保有效信息密度，减少时间开销。**

**\*注：“浅层DFS算法”是作者自己起的名字**

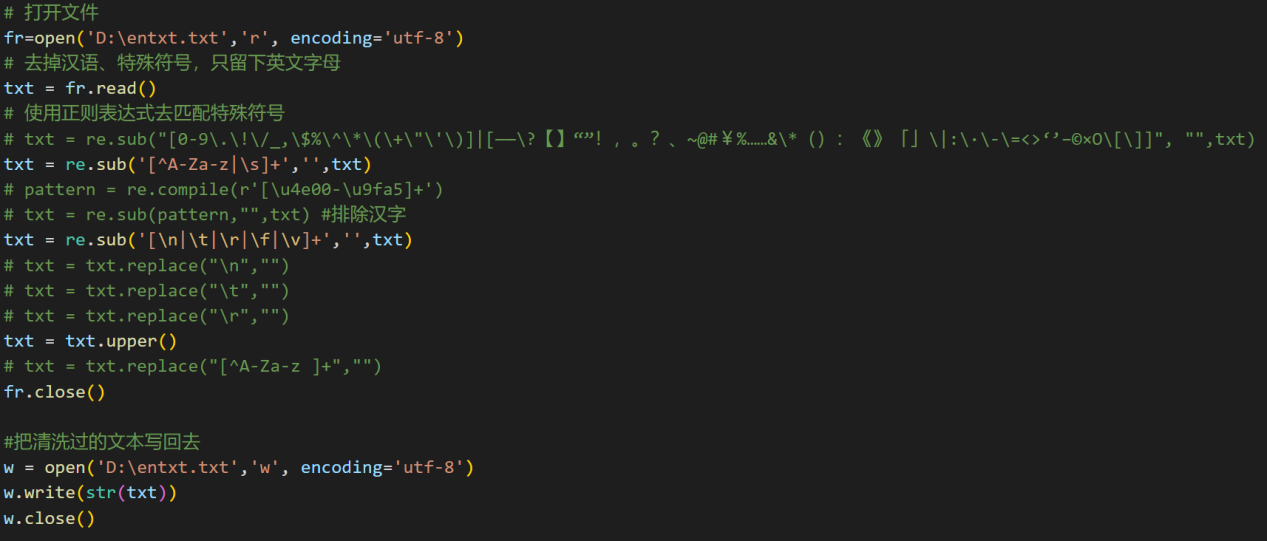
**三．数据清洗**

**将所有爬虫数据保存至txt文件中，随后进行数据清洗。**

**针对中文文本，利用正则表达式，结合汉字的GTF-8编码范围，去除除了汉字之外的所有字符，并写回文本文件：**

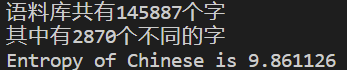


**针对英文文本，利用正则表达式，去除除了英文字母以及空格之外的所有字符，将小写字母全部转化为大写字母以便统计，并写回文本文件：**

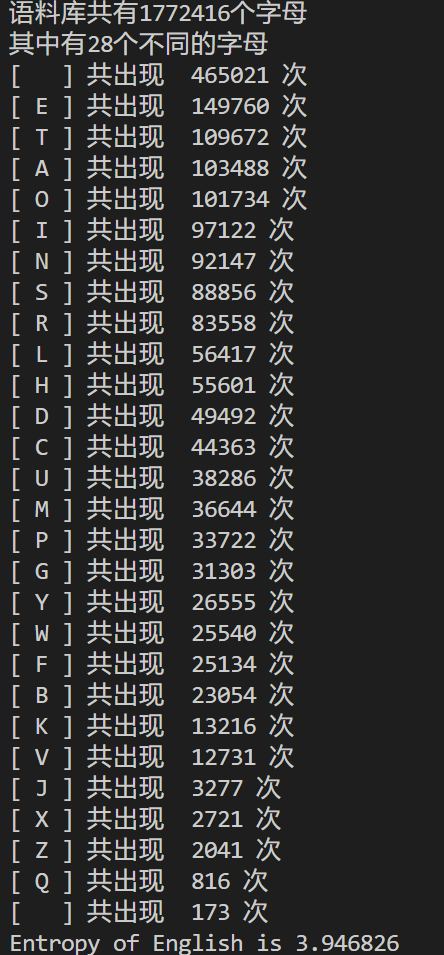


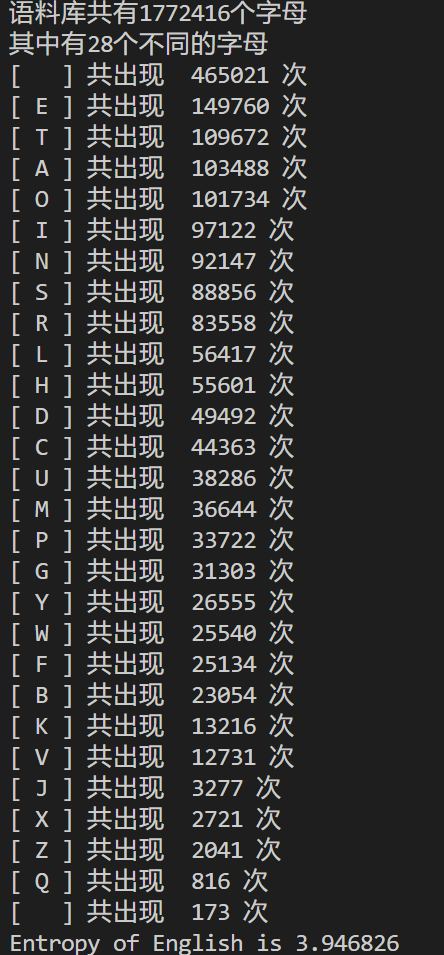
**四．字母/汉字出现特征统计及熵的计算**

**针对中文，我们首先选取max\_num=10，清洗后得到428KB语料，统计各个汉字出现频率，计算得到熵约为9.86bit：**

****

**针对英文，我们首先选取max\_num=10，清洗后得到2110KB语料，统计各个汉字出现频率，计算得到熵约为3.95bit：**

****

****

**\*注：英文文本字母总数不是27的原因是有特殊空字符的出现，这使得正则表达式无法排除它们。但它们出现频次极低，不影响计算结果。**

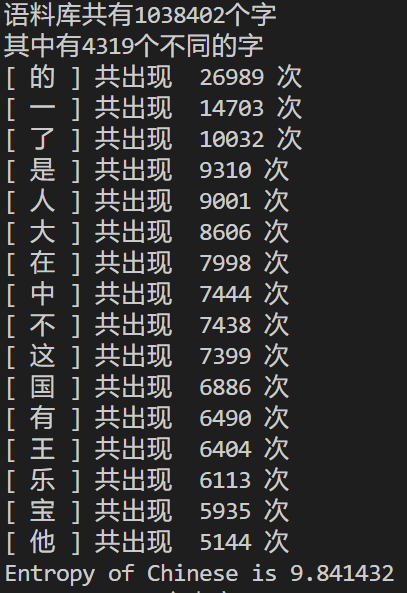
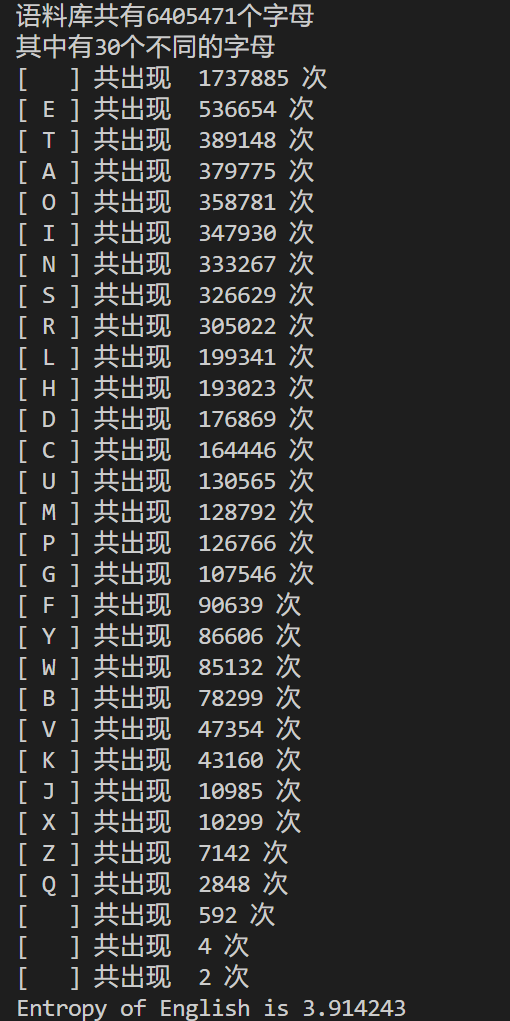
**计算高频汉字/字母出现频次如下：**

**下面我们扩大语料库并计算熵，观察熵的变化。针对语料库文本文件自身大小的精确调整是困难的，因为不同网页的信息含量各不相同，我们所能做的只是精确地改变max\_num，从搜索广度的角度完成语料库的逐步扩大。**

**针对中文，改变max\_num分别为50，100，200；针对英文，改变max\_num分别为20，30。统计熵的变化随语料库大小变化及趋势如下所示：**

**可以看出，扩大语料库之后中英文的熵值有所波动，逐渐向理论值靠近。在最大的语料库下，中英文熵值分别约为9.83bit和3.91bit。**

**分别对2994KB的中文语料库和7602KB的英文语料库进行字母/汉字频率统计，并与428KB的中文语料库和2110KB的英文语料库对比：**

** **

**可以看到，中英文扩充语料库前后的汉字/字母频次没有发生根本性的改变，整体格局保持一致，并且其频次数据很符合经验与直觉。但是英文空格有点太多了，这与PPT上结论不一致，推测是因为网站信息有许多自带冗余空格；此外，汉字“宝”出现频次过高，这应该是因为所爬虫的小说网站中主角名字里含有“宝”字。除此以外所得数据结论符合预期，计算得到的中英文熵与PPT结论相比差别很小，说明本研究数据具备一定正确性和代表性。**

**参考资料**

**[1]https://blog.csdn.net/weixin\_39525526/article/details/120502728**

**[2]https://blog.csdn.net/weixin\_54852327/article/details/115916146?utm\_medium=distribute.pc\_relevant.none-task-blog-2~default~baidujs\_baidulandingword~default-0-115916146-blog-118634078.235^v27^pc\_relevant\_3mothn\_strategy\_recovery&spm=1001.2101.3001.4242.1&utm\_relevant\_index=3**

**[3]https://cloud.tencent.com/developer/article/1478184**