自然语言处理第一次作业实验报告

朱辰 2022k8009970002

实验内容

网络爬虫：计算中英文本的熵

分别收集尽量多的英语和汉语文本，编写程序计算这些文本中英语字母和汉字的熵，对比本章课件第18页上表中给出的结果。然后逐步扩大文本规模，如每次增加固定的数量，如2M/5M等，重新计算文本规模扩大之后的熵，分析多次增加之后熵的变化情况。

爬虫

爬虫部分使用了 `requests` 和 `BeautifulSoup` 库来抓取网页数据。数据来源是人民网的中文与英文版 http://www.people.com.cn/和http://en.people.cn/。

具体代码见 `crawler.py` 文件。

核心流程分析

1. 整体流程：
   - 访问网页，获取网页内容
   - 解析网页内容，提取文本
   - 清洗文本，去除多余的空格和换行符
   - 将文本保存到文件中
   - 提取超链接并加入队列
2. 获取网页内容：
   使用了 `requests` 库的 `get` 方法来获取网页内容，并使用 `BeautifulSoup` 库解析HTML文档。

```
response = requests.get(url, headers=self.headers, timeout=10)
self.visited_urls.add(url)

if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser')
```

3. 解析网页内容：
   使用 BeautifulSoup 库解析HTML文档，提取文本内容。

```
    def extract_text(self, soup):
        """提取网页中的所有文本内容"""
        for script in soup(["script", "style"]):
            script.decompose()

        # 获取所有文本
        news = ''
        title = soup.title.text.strip() if soup.title else ''
        news += title
        for x in soup.find_all('div'):
            for y in x.find_all('p'):
                news += y.text.strip()

        return news
```

4. 清洗文本：
   使用正则表达式去除多余的空格和换行符。

```
def clean_text(self, text, keep_english=True, keep_chinese=True):
    """清洗文本，去除多余空格，并根据参数保留英文或中文"""
    text = re.sub(r'\s+', ' ', text).strip()

    result = ""
    if keep_english and keep_chinese:
        result = text
    elif keep_english:
        result = re.sub(r'[^a-zA-Z\s]', '', text)
    elif keep_chinese:
        result = re.sub(r'[^\u4e00-\u9fff\s]', '', text)

    # 输出清洗前后的文本长度
    print(f"清洗前文本长度: {len(text)}, 清洗后文本长度: {len(result)}")

    return result
```

5. 提取超链接并加入队列

```
def get_links(self, url, soup):
    """从网页中提取所有超链接并返回绝对URL"""
    links = []
    for link in soup.find_all('a', href=True):
        href = link['href']
        absolute_url = urllib.parse.urljoin(url, href)
        # 过滤非http链接
        if absolute_url.startswith('http'):
            links.append(absolute_url)
    return links
```

同时维护一个队列，存储待爬取的链接。每次从队列中取出一个链接进行爬取，直到达到设定的爬取数量。爬取完成后，将从该链接中提取的超链接加入队列中。

```
while queue and page_count < self.max_pages:
    ...
    url = queue.popleft()

    if url in self.visited_urls:
        continue

    ... #其它部分

    links = self.get_links(url, soup)
    for link in links:
        if link not in self.visited_urls:
            queue.append(link)

    page_count += 1
```

一些发现与思考

1. 黑名单设置

在爬取过程中，发现爬取网站中存在一些视频或者文档链接，这些链接并不包含文本内容。并且这些链接的内容通常较大，会严重拖慢爬取速度。为了避免这种情况，可以在爬取前设置一个黑名单，过滤掉这些链接。

```
blacklist = ['video', 'pdf', 'download', 'doc', 'xls', 'ppt', 'mp3', 'mp4']
# 如果url里含有被黑名单字符串，跳过
if any(black in url for black in blacklist):
    print(f"跳过黑名单URL: {url}")
    continue
```

2. 提取文本使用函数的选择

一开始，我直接使用了 BeautifulSoup 的 get_text() 方法来提取文本内容，但发现提取的文本并不完整，有时明明该网页有内容却提取不到。经过调试，发现 get_text() 方法会忽略一些标签内的文本内容，因此我决定使用ppt中的方法手动选择标签。

3. 无效链接的处理

经过对爬虫部分的观察，发现由于在爬取时对所有链接都进行了爬取，导致爬取的链接有大量不符合要求的链接，如人民网以外的站点和人民网其它语言的站点，猜测来源于类似“友情链接”之类的超链接或者新闻里为了引用而添加的链接。

因此仿照黑名单机制，设置一个白名单，只爬取人民网对应语言的链接。同时对欸名单进行了一定的扩展，增加了其它语言的部分

```
blacklist = ['video', 'pdf', 'download', 'doc', 'xls', 'ppt', 'mp3', 'mp4','swahili','italian'
# 如果url里含有被黑名单字符串，跳过
if any(black in url for black in blacklist):
    print(f"跳过黑名单URL: {url}")
    continue

if not keep_english:
    whitelist =['people']
else:
    whitelist = ['en.people']
# 如果url里不含有白名单字符串，跳过
if not any(white in url for white in whitelist):
    print(f"跳过非白名单URL: {url}")
    continue
```

熵的计算

这一部分使用 collections.Counter 来统计字符频率，然后计算熵。使用 pandas 和 matplotlib 来可视化熵的变化。

出于方便考虑，我并没有采用改变爬取文本大小的方式来控制文本大小，而是直接爬取大量文本，然后在数据处理时进行分割。

具体代码见 `analyze.py` 文件。

核心流程分析

1. 整体流程：
    - 读取文本文件
    - 选择文本量
    - 统计字符频率
    - 计算熵
    - 可视化熵的变化
2. 统计字符频率与计算熵：

```
def calculate_entropy(text):
    # 计算文本熵
    counter = Counter(text)
    length = len(text)
    probabilities = [count / length for count in counter.values()]
    entropy = -sum(p * math.log2(p) for p in probabilities)
    return entropy
```

3. 控制文本量：

```
# 计算不同文本量下的熵
results = []
for i in range(step_size, len(text) + 1, step_size):
    sample = text[:i]
    size_mb = i / (1024 * 1024)
    entropy = calculate_entropy(sample)
    results.append({'Size (MB)': round(size_mb, 2), 'Entropy (bits)': round(entropy, 4)})

# 如果最后一个样本不是整数倍step_size
if len(text) % step_size != 0:
    sample = text
    size_mb = len(text) / (1024 * 1024)
    entropy = calculate_entropy(sample)
    results.append({'Size (MB)': round(size_mb, 2), 'Entropy (bits)': round(entropy, 4)})
```

结果与分析

1. 文本爬取结果

通过控制 `max_pages` 参数,设置其值为中文15000英文10000，即对于每个语言爬取10000个网页。爬取完成后，分别得到了中文和英文的文本数据，大小分别为：

- 中文文本：约6.86 MB
- 英文文本：约47.9 MB

爬取内容展示：

- 中文

史地落到中国无产阶级身上在民主和科学两面大旗的指引下新文化运动成为空前深刻的思想解放运动年以先进青年知识分子为先锋的五四运动爆发中国工人阶级作为独立的政治力量登上历史舞台使运动发展成广大人民群众参加的彻底反帝反封建的伟大爱国革命运动五四运动的胜利激发起中国人民和中华民族追求真理追求进步的伟大觉醒推动了社会主义思潮在中国的蓬勃兴起开启了中国新民主主义革命的伟大征程五四运动后新思潮大量涌现研究和宣传社会主义逐步成为中国进步思想界的主流随着俄国十月革命的影响渐次扩大以李大钊为代表的先进分子开始在中国比较系统地传播马克思主义经过反复的比较推求越来越多的进步青年被马克思主义高度的科学性和革命性吸引从民主主义者转变为马克思主义者为无产阶级政党的创建准备了思想条件和干部条件年月出版的陈望道译共产党宣言年月张亮译列宁传列宁全书第四种年月田诚著共产主义与智识阶级在广泛传播马克思主义宣传劳工运动的过程中陈独秀和李大钊开始酝酿和准备建立共产党组织年在中国工人阶级最密集的中心城市上海以上海马克思主义研究会为基础中国的第一个共产党早期组织正式成立上海的共产党早期组织积极推动各地共产党早期组织的建立成为各地共产主义者进行建党活动的联络中心实际上起着中国共产党的发起组织的作用在上海及北京共产党早期组织的联络和推动下年秋至年春武汉长沙广州济南等地的先进分子以及旅日旅法留学生和华侨中的先进分子相继建立了共产党早期组织各地共产党早期组织成立后有组织有计划地开展各项工作进一步扩大了马克思主义的影响促进了马克思主义同中国工人运动的结合为建立全国统一的中国共产党夯实基础围绕建立一个什么样的党和怎样建党等关键而重要的问题中国早期马克思主义者进行了积极的探索初步明确了党的根本性质奋斗目标组织原则和革命手段为建立新型的无产阶级政党筑牢理论根基同改良主义无政府主义等反马克思主义思潮的论战帮助一批进步青年划清科学社会主义与资产阶级小资产阶级社会主义流派的界限走上马克思主义的道路年秋李大钊用的英文打字机年月袁让译工钱劳动与资本年月太柳译共产党底计划康民尼斯特丛书第一种年月共产国际代表抵达上海与上海共产党早期组织成员取得联系经讨论并征求陈独秀和李大钊的意见决定在上海召开中国共产党第一次全国代表大会上海共产党早期组织负责筹备会议

- 英文

optionsIn pics Stage at th Tour of Hainan cycling raceChinese FM meets Thai princess SirindhornChinese FM meets Thai princess SirindhornChinese FM meets Thai princess SirindhornYouth talk The cultural bonds between China and Southeast AsiaA feast for the senses in Heze Russian visitor savors a Peony BanquetStunning scenery of Emerald Lake in Mangya NW Chinas QinghaiThe Foreigners Fortune Quest Changle Lianjiang Fresh seafood paradiseTrending in China Luoyangs blooming peony paradiseYouth talk The cultural bonds between China and Southeast AsiaA feast for the senses in Heze Russian visitor savors a Peony BanquetStunning scenery of Emerald Lake in Mangya NW Chinas QinghaiThe Foreigners Fortune Quest Changle Lianjiang Fresh seafood paradiseTrending in China Luoyangs blooming peony paradiseYouth talk The cultural bonds between China and Southeast AsiaTrumps tariffs spark bitter aftertaste among Italian wine producersConsumption promotion campaign for foreigners set to launchChinaLaos Railway transports over crossborder passengers in yearsChinas braincomputer interface innovation in fast laneY antisubmarine patrol aircraft takes off for trainingPreliminary round of th Chinese Bridge competition held in RussiaUnderstand ChinaTrend TrackerCity WalkCartoonThe Cultural Silk RoadCalendar for Traditional Chinese CultureOur China StoriesYouth in the New EraTrumps tariffs spark bitter aftertaste among Italian wine producersConsumption promotion campaign for foreigners set to launchChinaLaos Railway transports over crossborder passengers in yearsChinas braincomputer interface innovation in fast laneY antisubmarine patrol aircraft takes off for trainingPreliminary round of th Chinese Bridge competition held in RussiaTrumps tariffs spark bitter aftertaste among Italian wine producersConsumption promotion campaign for foreigners set to launchTrumps tariffs spark bitter aftertaste among Italian wine producersTrumps tariffs spark bitter aftertaste among Italian wine producersConsumption promotion campaign for foreigners set to launchConsumption promotion campaign for foreigners set to launchConsumption promotion campaign for foreigners set to launchChinaLaos Railway

- 运行时截图

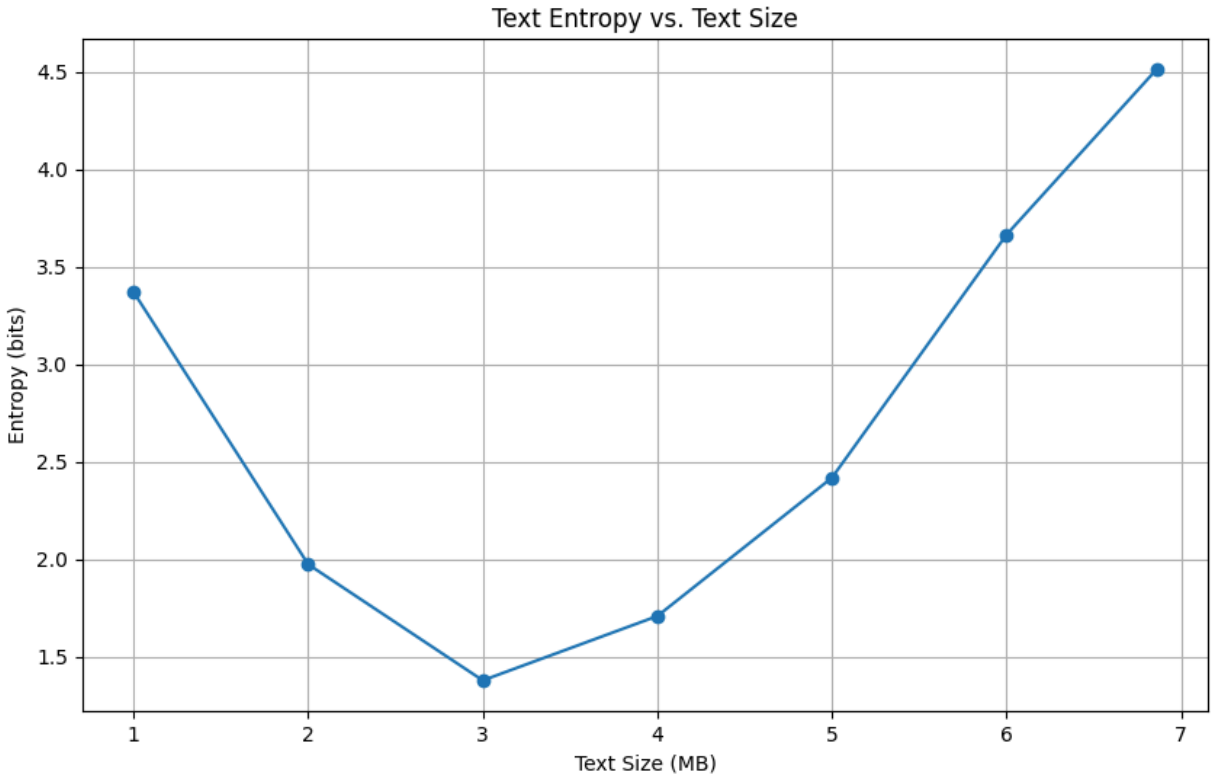
```
正在爬取: http://en.people.cn/n3/2021/1102/c90000-9914261.html 已爬取 2836/10000 ,已爬取文件大小: 11336848 bytes
清洗前文本长度: 17162, 清洗后文本长度: 16623
成功写入16623个字符到文件
正在爬取: http://en.people.cn/n3/2021/1029/c90000-9913342.html 已爬取 2837/10000 ,已爬取文件大小: 11359963 bytes
清洗前文本长度: 483, 清洗后文本长度: 444
成功写入444个字符到文件
正在爬取: http://en.people.cn/n3/2021/1029/c90000-9913291.html 已爬取 2838/10000 ,已爬取文件大小: 11359963 bytes
清洗前文本长度: 2711, 清洗后文本长度: 2616
```

具体爬取的文本数据存储在 `result_CH.txt` 和 `result_EN.txt` 文件中。

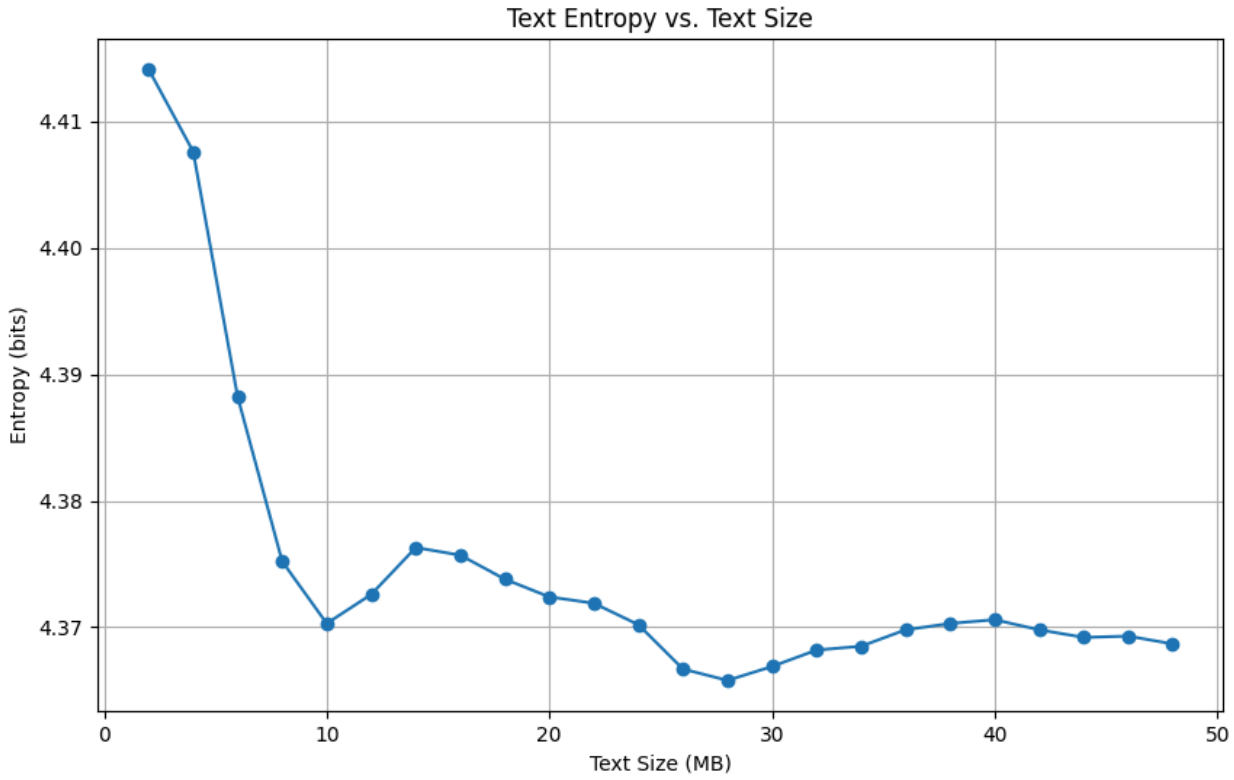
2. 熵的计算结果

通过对爬取的文本进行熵的计算，得到了不同文本量下的熵值。

中文：



英文：



最终在最大文本量下，中文和英文的熵值分别为：

- 中文：4.5117 bits
- 英文：4.3687 bits

与文献结果相比，中文的熵值低于课件中的结果，英文的熵值基本相同

具体结果见 `analysis_EN/` 和 `analysis_CH/` 目录中内容

结果分析

对于中文，随着文本量的增加，熵值仍在增加，且波动幅度较大

对于中文文本熵值明显小于文献值的猜测：

1. 爬取数据量较小，可以看到，随着数据量增加，熵值仍在增加，若继续增加数据量，熵值可能会接近文献值
2. 文本来源于人民网，人民网的文本内容相对单一，可能导致熵值偏低，比如经过观察，人民网报道中一些常用的词语如“人民”、“中国”等出现频率较高，导致熵值偏低
3. 数据未清晰干净，存在例如“年/月/日”等非正文内容，可能导致熵值偏低

值得注意的是，尽管爬取中文的网页数大于英文，但中文文本的大小却显著小于英文文本的大小，这一点可能是由于所爬取网站（人民网）的内容差异造成的。

对于英文，随着文本量的增加，熵值逐渐趋于稳定，且波动幅度始终不大，说明英文文本的字符分布相对均匀。