自然语言处理 第一次作业*

蔡宇哲
2022E8014682046
Institute of Automation

1 数据准备

1.1 中文语料

中文语料为从乐乐课堂爬取的 29731 篇高中作文。原始数据中含有少量的乱码和非汉字符号，同时为只计算汉字的熵，需将预料中的标点符号去除。数据经清洗后共 69MB。

1.1.1 数据清洗

\u4e00和\u9fa5是中文 unicode 编码的开始和结束，使用其对数据进行清洗。

```
import re
patten = re.compile(r"[^\u4e00-\u9fa5 ]+", re.UNICODE)
cleaned_corpus = patten.sub('', raw_corpus)
```

1.1.2 中文语料示例

- 清洗前

  感动的心藏在槐树下。氤氲的夜色浸染着紫色的窗枝，梦里似乎又闯到槐花淡淡的幽香似一道澄清的柔波……月下。小路。老槐树……又是一个有月的晚上，风渐起渐落，吹乱了人们的头发，吹乱了人们的思绪，只有那颗心显得格外温暖。槐树在小路的尽头，树下有间馄饨店，摊主是一位慈祥的老伯，六一的样子。虽然在小路尽头，但生意格外好。

- 清洗后

  感动的心藏在槐树下氤氲的夜色浸染着紫色的窗枝梦里似乎又闯到槐花淡淡的幽香似一道澄清的柔波月下小路老槐树又是一个有月的晚上风渐起渐落吹乱了人们的头发吹乱了人们的思绪只有那颗心显得格外温暖槐树在小路的尽头树下有间馄饨店摊主是一位慈祥的老伯六一的样子虽然在小路尽头但生意格外好

*Data and code: https://github.com/Zhe0311/entropy_of_language

1.2 英文语料

英文语料为从 Wikipedia 上爬取的对英文词语的解释，英文词语来自于WikiCatSum数据集。由于该数据集数据量较大，只选用其`train.src`中的部分词语，之后爬取 Wikipedia 对于所有词语的释义。

英文语料包括 259255 条 Wikipedia 对词语的 Description，共约 82MB。

1.2.1 数据清洗

原始数据中存在较多的乱码、拉丁文等非英文字母符号，为计算英语语料中英文字母的熵，对于英文语料，只保留英文字母（大写和小写）。

```
import re
cleaned_corpus = [''.join(re.findall(r'[A-Za-z]', text)) for text in raw_corpus]
```

1.2.2 英文语料示例

- 清洗前

  ```
  Carmarthenshire (; Welsh: Sir Gaerfyrddin; [sir gar vərðn] or informally Sir
  Gâr) is a unitary authority in southwest Wales, and one of the historic
  counties of Wales. The three largest towns are Llanelli, Carmarthen and
  Ammanford. Carmarthen is the county town and administrative centre.
  ```

- 清洗后

  ```
  CarmarthenshireWelshSirGaerfyrddinsirgrvrnorinformallySirGrisaunitaryauthori
  tyinsouthwestWalesandoneofthehistoriccountiesofWalesThethreelargesttownsareL
  lanelliCarmarthenandAmmanfordCarmarthenisthecountytownandadministrativecentre
  ```

2 汉字/字母概率计算

2.1 建立词表

对于清洗后的中文和英文语料，首先进行词表的建立，词表形式如下

```
{
    token1: number1,
    token2: number2,
    ...
}
```

token 为汉字（字母），number 为该 token 在语料中出现的次数，即统计每一个汉字（字母）在中文（英文）语料中出现的次数。

2.2 汉字/字母概率计算

根据词表即可计算得到汉字/字母出现的频率，由此作为概率。具体做法为将词表中每个 token 的频数除以此表中所有 token 频数之和。

中文语料（69MB）中汉字的概率分布[1]：

```
{
    '的': 0.053635171766986720,
    '我': 0.023400041489229043,
    '一': 0.019817753119813066,
    '是': 0.017001943779519225,
    '了': 0.014815700674071201,
    '不': 0.014139984525863434,
    '在': 0.011719751805682861,
    '人': 0.010754466759143726,
    '有': 0.010431001974850748,
    '们': 0.009226908780770798,
    ...
}
```

英文语料（82MB）中字母的概率分布[2]：

```
{
    'E': 0.11280347933278792
    'A': 0.09551479730185844
    'I': 0.08203709113202019
    'T': 0.07810749886954046
    'N': 0.07425067574369493
    'S': 0.07127359485376289
    'O': 0.07054930699615834
    'R': 0.06605005687510414
    'L': 0.04450638830346183
    'H': 0.04374229470353272
    'D': 0.03786952271219087
    'C': 0.034764851463577065
    'M': 0.027630060787843733
    'U': 0.02749578118015594
    'F': 0.022777517038388706
    'P': 0.02120759269332389
    'G': 0.018842919616001305
    'B': 0.017397511872592607
    'Y': 0.015859561666702705
    'W': 0.014800882019878094
    'V': 0.008562815793203846
    'K': 0.007518480345832289
```

[1]由于中文词表较大，此处仅展示概率由大到小排列前 10 的汉字

[2]大、小写字母均按大写字母统计

```
    'J': 0.002106324388158654
    'X': 0.001823291428014754
    'Z': 0.0016410093926006448
    'Q': 0.000866693489613033
}
```

3 汉字/字母熵的计算

3.1 中文语料

对于 69MB 的中文语料，其熵

$$\text{Entropy}_{\text{CHS}} = -\sum_{c \in \text{vocab}} p(c) \log p(c) = 9.339044$$

课件中中文的熵约为 9.71，略大于本实验得到的数值，分析原因可能如下：

1. 本实验所用的语料均为高中作文，领域较单一，可能会导致熵减小；
2. 实验所用数据量不够大，熵仍未收敛，4.1中可以看到随着语料规模的扩大，熵的值有下降的趋势。

3.2 英文语料

对于 82MB 的英文语料，其熵

$$\text{Entropy}_{\text{EN}} = -\sum_{c \in \text{vocab}} p(c) \log p(c) = 4.178977$$

课件中中文的熵约为 4.03，略小于本实验得到的数值，二者的差距较小，可能是因为本实验所用的英文语料包含了众多领域，模式比较全面，更能代表总体分布。

4 熵随文本规模的变化

4.1 中文语料

中文语料由 29731 篇高中作文构成，约 69MB，平均意义下，每 849 篇作文大小约为 2MB。在实验中，每次增加 849 篇作文（即大约增加 2MB 文本），计算熵的值。如图1所示是汉字熵随文本规模变化的变化情况。

4.2 英文语料

英文语料由 259255 条 Wikipedia 对词语的 Description 构成，约 82MB，平均意义下，每 6323 条文本大小约为 2MB。在实验中，每次增加 6323 条文本（即大约增加 2MB 文本），计算熵的值。如图2所示是字母熵随文本规模变化的变化情况。

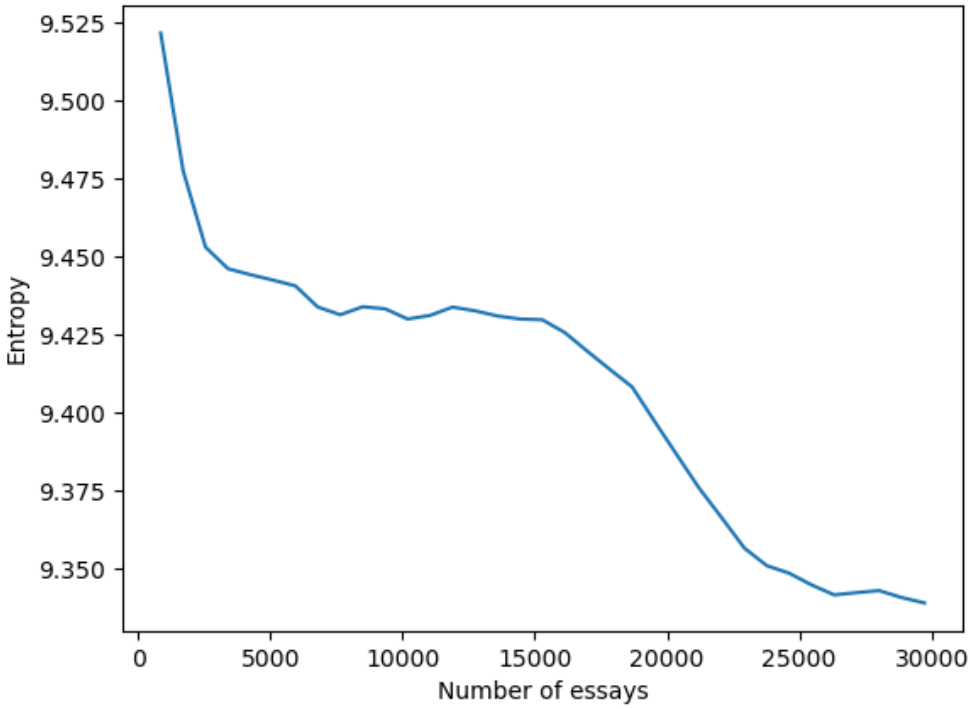

图 1: 汉字的熵与文本规模之间的关系曲线

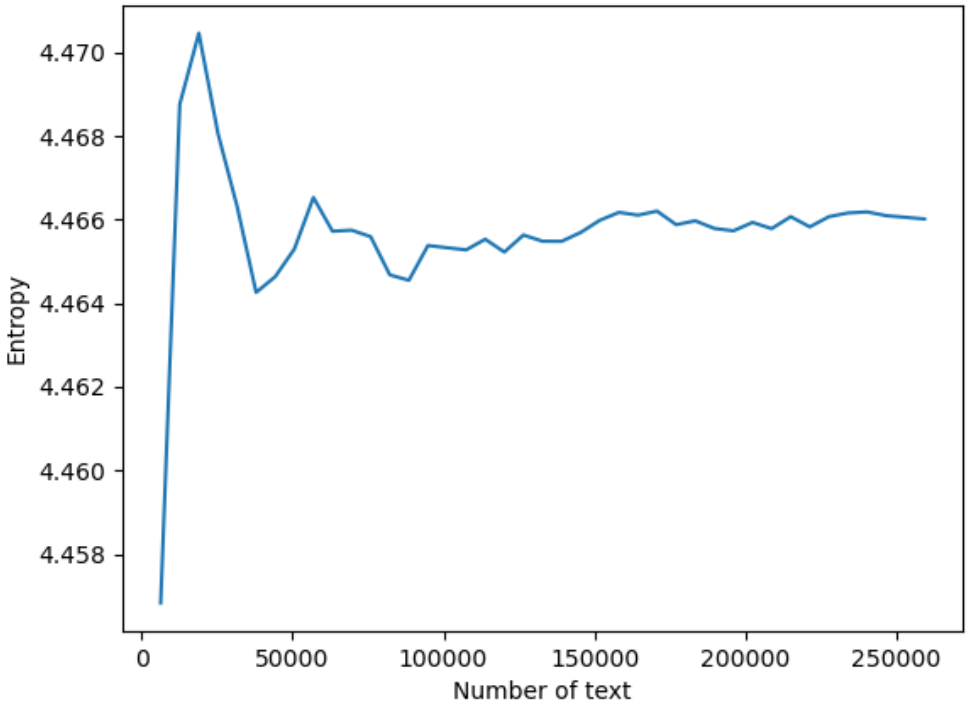

图 2: 字母的熵与文本规模之间的关系曲线

4.3 对比与分析

可以发现对于中文汉字，总体来看熵随着语料规模的增大而减小；而对于英文字母，总体来看熵随着语料规模的增大先波动后逐渐稳定在某一个值附近。

汉字与字母的一大区别是汉字的取值空间远大于字母的取值空间。随着中文语料规模的增大，汉字的词表不断扩大，随机性的影响逐渐减小，常用字的占比不断增大，熵呈现单调递减的规律；对于字母，由于其只有 26 个取值，在语料规模较小时，随机性对于熵的影响更大，所以在语料规模较小时熵会出现振荡，后随着语料规模的增大而逐渐收敛。

二者相同之处在于，随着语料规模的增大，其统计的概率分布都更加接近真实概率分布，熵的值逐渐收敛到真实值。

5 总结

本报告中首先介绍了中文语料和英文语料的组成、爬取与清洗；其次介绍了汉字、字母的概率分布和熵的计算；最后探究了汉字、字母的熵随着语料规模变化的变化情况，即，总体来看，汉字熵随着语料规模的增大而减小，字母熵随着语料规模的增大先波动后逐渐稳定在某一个值附近。