形状, 圆圈

描述已自动生成

**卡通人物

中度可信度描述已自动生成**

**深度学习与自然语言处理第一次作业**

中文信息熵的计算

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 专业名称 | 电子信息 |
| 学生姓名 | 彭程 |
| 学号 | ZY2103807 |
| 指导老师 | 秦曾昌 |

2022年 4月

# 实验背景

## 问题描述

通过阅读参考文献 *An Estimate of an Upper Bound for the Entropy of English*[1],学习信息熵的计算方法。利用此方法计算16本中文小说的中文信息熵并分析

## 实验原理

**1.2.1 信息熵定义**

1948年，香农从热力学当中借鉴提出信息熵的概念，解决了对信息量化度量的问题。其定义为：

**1.2.2 统计语言模型**

假定*S*表示某个有意义的句子，由一连串特定顺序排列的词，，，…. 组成。其中*n*代表句子的长度。现在我们想知道*S*在文本中出现的可能性，即：

利用条件概率公式：

当计算，仅存在一个参数；计算，存在两个参数，此后越往后累积的，计算难度加大。所以马尔可夫提出一种假设：假设出现的概率只与前面N-1个词相关，当N=2时，就是二元模型，N=3就是三元模型，本次实验分别使用一元模型、二元模型，三元模型来统计语料库字数，分词个数，平均词长，模型长度，基于模型的中文信息熵和运行时间。

# 实验方法

## 实验过程

1. 下载数据集并进行预处理，删除所有的隐藏符号，标点符号，以及非中文字符；
2. 分别基于词和基于字的一元模型、二元模型、三元模型计算中文信息熵；

## 代码设计

在本次实验中，使用Python语言完成数据处理和算法设计。详细代码见.py文件。代码文件包括主程序文件、数据读取及处理文件、信息熵计算问价。

# 实验结果与分析

## 基于字的实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分词模型 | 语料字数 | 平均词长 | 信息熵（比特/词） | 运行时长(秒) |
| Unigram | 7420145 | 1 | 9.50 | 2.85 |
| Bigram | 7420145 | 1 | 6.69 | 8.5 |
| Trigram | 7420145 | 1 | 23.94 | 21.74 |

表 1. 基于字的运行结果

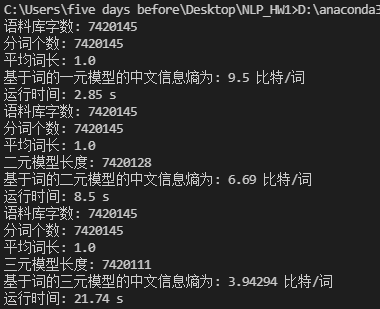


图1. 基于字的运行结果

## 基于词的实验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 分词模型 | 语料字数 | 分词个数 | 平均词长 | 信息熵（比特/词） | 运行时长(秒) |
| Unigram | 7420145 | 4430791 | 1.67 | 12.01 | 42.86 |
| Bigram | 7420145 | 4430791 | 1.67 | 6.89 | 46.87 |
| Trigram | 7420145 | 4430791 | 1.67 | 2.42 | 59.13 |

表 2. 基于词的运行结果

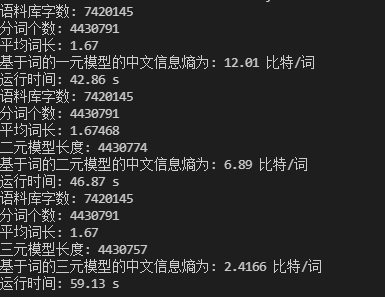


图2. 基于词的运行结果

## 结果分析

通过以上实验可以看出分词所占用的时间较长，而当n增大时，程序运行时间也会变长。

**附录：**

**参考文献**

[1] Brown, Peter & Della Pietra, Stephen & Pietra, Vincent & Lai, Jenifer & Mercer, Robert. (1992). An Estimate of an Upper Bound for the Entropy of English. Computational Linguistics. 18. 31-40.

[2] https://blog.csdn.net/weixin\_42663984/article/details/115718241

[3] https://blog.csdn.net/qq\_40412713/article/details/115742092