Course_Project_1

专业班级：**提高2201班**

姓名：　　**王翎羽**

学号：　　**U202213806**

实验工具

PyCharm 2023.3

依赖项: **Jupyter Notebook**、**python-docx**

实验结果

英文

对于`Elon_Musk_Speech.docx`

```
Order 0 Markov Model, Entropy: 4.089509713537937
Order 1 Markov Model, Entropy: 3.26347630607172
Order 2 Markov Model, Entropy: 2.2088364510658356
Order 3 Markov Model, Entropy: 1.2720369608559907
Order 4 Markov Model, Entropy: 0.7359939792669795
Order 5 Markov Model, Entropy: 0.4698597610678271
Order 6 Markov Model, Entropy: 0.29026720582504056
Order 7 Markov Model, Entropy: 0.17660436174822536
Order 8 Markov Model, Entropy: 0.11286164678247694
Order 9 Markov Model, Entropy: 0.0711662577925939
```

对于这个结果，可以认为0-1阶的Markov模型近似满足真实的英文熵值。
而由于所提供的英文文本内容有限，所以对于更高阶的Markov模型，得到的熵值不会趋于值约为1.4的极限。

转移概率的结果已保存为文本文件，见附件。

于是我从**美国当代英语语料库**中下载了含有10亿单词以上的英文语料文件，喂给模型计算。结果如下:

```
Order 0 Markov Model, Entropy: 4.254037364372333
Order 1 Markov Model, Entropy: 3.5462728325582686
Order 2 Markov Model, Entropy: 2.9574901930307393
Order 3 Markov Model, Entropy: 2.3822194443288214
Order 4 Markov Model, Entropy: 2.0052218847816046
Order 5 Markov Model, Entropy: 1.7797140240443352
Order 6 Markov Model, Entropy: 1.606931938438108
Order 7 Markov Model, Entropy: 1.4367615117073607
Order 8 Markov Model, Entropy: 1.2568568346679978
Order 9 Markov Model, Entropy: 1.0718830859964419
```

也许是数据的数量还不够，或者是预处理数据的问题，随着阶数的升高，熵值并没有收敛在1.4.....

中文

对于`Zhengfei_Ren_Email.docx`

```
Order 0 Markov Model, Entropy: 8.295259531645113
Order 1 Markov Model, Entropy: 2.1807726849587
Order 2 Markov Model, Entropy: 0.5487715179179455
Order 3 Markov Model, Entropy: 0.11718543421670218
Order 4 Markov Model, Entropy: 0.04765213263572899
```

由于数据过少，导致只有零阶熵的数据比较准确。

转移概率的结果已保存为文本文件，见附件。

> 参考文献：（其实是从知乎上找到hhhh）
> 在国内最早冯志伟先生用了将近 10 年的时间，进行手工查频，从小到大地逐步扩大统计的规模，建立了 6 个不同容量的汉字频度表，最后根据这些不同的汉字频度表，逐步地扩大汉字的容量，终于在 70 年代末期首次计算出了在不考虑上下文影响的前提下汉字信息熵的值是 9.65 比特.....黄萱菁等在 4 年的《人民日报》语料的基础上，所求得的零阶熵、一阶熵、二阶熵分别为 9.62，6.18 和 4.89 比特。 刘源给出汉字熵的计算结果是9.71 比特。 孙帆等基于词的语言模型估计方法比基于字的直接计算方法得到了汉字熵的更为精确的估计，其熵值为 5.31 比特。

于是我在Github上找到了2015-2019年每天新闻联播的语料，数据的数量级在千万级别，喂给模型计算。得到的结果更接近于文献中的数据。结果如下。

```
Order 0 Markov Model, Entropy: 9.499217621115344
Order 1 Markov Model, Entropy: 5.817780770339022
Order 2 Markov Model, Entropy: 3.30672556299714
Order 3 Markov Model, Entropy: 1.5715582554874472
Order 4 Markov Model, Entropy: 0.8586357082190496
```

One More Thing

我做了点有意思的事情，就是让Markov模型生成了200字的文本。结果如下。

```
print(generate_text(Chinese_Text_model, 3, 200))
```

在 86ms 的 2024.05.22 17:28:47 执行

律和政策环境三是增进友好感情立法机关应在促进双边合作中国常驻联合国代表贾法里日表示敌对势力近期在伊朗煽动的骚乱已经得到平息伊朗伊斯兰共和国总统施泰因迈尔表示特朗普似乎把惩罚性关税当作用来对付所有人和事的万能药美国单方面宣布退出与相关国家的投资贸易制度体系深化金融开放创新到加快政府职能转变财税金融国企国资对外开放生态文明等关键领域取得重大成就中国共产党第十九届中央委员会工作报告审议通过了全国人大宪法

喂了新闻联播的模型一张口就是国际形势hhhh

鸣谢

太感谢GPT-4o啦！不仅帮我写导入数据的代码，还帮我改了半个小时没找出来的Bug.