NLP 大作业——金庸小说信息熵计算

学院： 自动化科学与电气工程学院 **姓名：**王明贤 **学号：**ZY2103526

一、前期准备

1.信息熵

熵的概念由德国物理学家克劳修斯提出，泛指某些物质系统状态的一种量度，某些物质系统状态可能出现的程度，熵的本质是一个系统“内在的混乱程度”。信息论开创者香农指出信息的信息量与其不确定度有直接关系。

根据信息量的性质构造信息函数：$I(x)=-\log p(x)$，而信息函数的数学期望为信息熵 $H(x)=-\sum_{x\in X} p(x)\log p(x)$。

对于两个随机变量 X 和 Y，在 X 已知的前提下 Y 的熵定义为 Y 的信息熵，即 $H(Y\mid X)=-\sum_{x_i,y_i}^{m,n} p(x_i,y_i)\log p(y_i\mid x_i)$。

2.语言模型

语言模型是用于评估文本序列符合人类语言使用习惯程度的模型。统计语言模型是基于预先人为收集的大规模语料数据，以真实的人类语言为标准，预测文本序列在语料库中可能出现的概率，并以此概率去判断文本是否“合法”，是否能被人所理解。

若认为前后语义独立，以链式法则展开联合概率，将产生很多问题。为避免参数空间过大，以及数据稀疏严重的问题，引入马尔可夫假设，随意一个词出现的概率只与它前面出现的有限的一个或者几个词有关。由此产生了N-Gram模型。

本文中考虑一个词与其前两个词有关，采用的是 Tri-Grams 模型。即 $p(S)=p(w_1w_2\cdots w_n)=p(w_1)p(w_2\mid w_1)\cdots p(w_n\mid w_{n-1}w_{n-2})$。

二、问题描述与分析

1.问题描述：

给定文本数据库，链接：https://share.weiyun.com/5zGPyJX，首先阅读文章：Entropy of English，然后在三元模型基础上计算中文(分别以词和字为单位)的平均信息熵。

2.问题分析

对于三元模型，需要计算条件信息熵 $H(Y|X)=-\sum_{x_i,y_i}^{m,n} p(x_i,y_i)\log p(y_i|x_i)$，其中的每一项条件概率需要通过极大似然估计得到，即对语料库中频数进行统计。

$$p(w_n|w_{n-1}w_{n-2})=\frac{C(w_{n-2}w_{n-1}w_n)}{C(w_{n-2}w_{n-1})}$$

因此需要对语料库进行分词，jieba 分词是 python 中文分词常用的分词库，可以根据不同应用需求进行分词。同时对数据需进行预处理，方便后续使用。

三、算法设计

该问题是一个自然语言处理的基础问题，利用信息熵定义结合语言模型对词频统计计算文本信息熵。

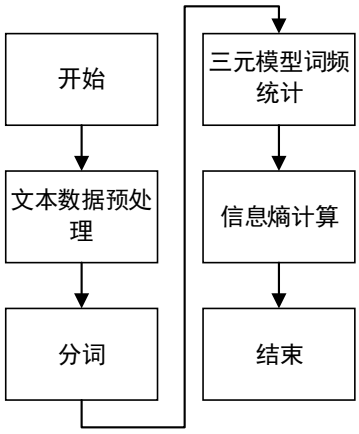

图 1

如图 1 所示，该算法分为以下 4 步：

（1）文本数据预处理：包括文本读取、去除特殊字符等；

（2）分词：包括 jieba 分词和按字分词；

（3）三元模型词频统计：根据模模型分别对联合概率和条件概率进行极大似然估计；

（4）信息熵计算：根据信息熵计算每类词段信息熵并求和。

注：本程序实现参考了 CSDN 博客：深度学习与自然语言处理实验——中文信息熵的计算。对语料库生成及预处理进行了调整改进，并增加了按字分词情况下信息熵计算。

四、运行结果

1.运行结果

（1）以字为单位计算平均信息熵

语料库字数	7266523
分词个数	7266523
平均词长	1.0
三元模型长度	7266491
基于三元模型的信息熵	3.94781 比特/词
运行时间	15.79861s

（2）以词为单位计算平均信息熵

语料库字数	7266523
分词个数	4273193
平均词长	1.70049
三元模型长度	4273161
基于三元模型的信息熵	2.29768
运行时间	39.5937s

2. 结果分析

以词为单位进行信息熵计算由于需要进行分词显然运行时间更长，且平均词长更长。由于词包含的内容更丰富，直观上看其信息熵较按字分词小。

五、总结体会

对自然语言的处理人们也进行了很多的思考，从开始的建立语法规则对语言进行描述，到现在以统计、信息论为依据。语言模型很好地刻画了语言本身的特点，传播信息的多少又恰能通过信息熵度量。