中文平均信息熵的计算

李志渊
PT2200081

一．问题描述

根据 16 本金庸小说，在排除标点符号及停词的前提下，计算其中文信息熵。

二．具体实现方法

2.1 资料预处理

由于一元模型不需要考虑上下文关系，所以与二元、三元模型不同，直接将 txt 文件合并成一个文件，通过 jieba 分词，得到所需要的 txt 格式语料库。

二元模型和三元模型需要考虑上下文，通过中文停词表清理初始资料，生成新的 txt 格式语料库。

2.2 词频

一元模型只需要统计每个词在语料库中出现的次数，以获得词频表。二元模型需要统计每个词在语料库中出现的次数，以获得词频表，在计算条件概率$P(\mathrm{w_i} \mid w_{i-1})$时作为分母，并且需要统计每个二元词组在语料库中的出现次数，以获得二元模型词频表。三元模型需要统计语料库中每个二元词组出现的频率，得到二元模型词频表，在计算条件概率$P(\mathrm{w_i} \mid w_{i-2}, w_{i-1})$时作为分母，同时统计语料库中每个三元词组的出现频率，得到三元模型词频表。

2.3 信息熵

一元模型的信息熵计算公式为

$$H(X) = -\sum_{x\epsilon X, y\epsilon Y} P(x) log P(x)$$

其中 P(x)近似等于每个词在语料库中出现的频率

二元模型的信息熵计算公式为

$$H(X|Y) = -\sum_{x\epsilon X, y\epsilon Y} P(x,y) log P(x|y)$$

其中联合概率 P(x,y)可近似等于每个词在语料库中出现的频率，条件概率$P(x|y)$可以近似等于每个二元词组出现的次数与其第一个词为词首的二元词组的次数的比值。

三元模型的信息熵计算公式为

$$H(X|Y,Z) = -\sum_{x\epsilon X, y\epsilon Y, z\epsilon Z} P(x,y,z) log P(x|y,z)$$

其中联合概率$P(x,y,z)$可近似等于每个三元词组在语料库中出现的频率，条件概率$P(x|y,z)$可近似等于每个三元词组出现的次数与其前两个字为词首的三元词组的次数的比

值。

三．运行结果

表 1 各书平均信息熵

书名	1-gram	2-gram	3-gram
三十三剑客	11.654	2.9647	0.2879
书剑恩仇录	11.7007	5.0327	1.0507
侠客行	11.1696	4.9674	1.1364
倚天屠龙记	11.7391	5.5257	1.3419
天龙八部	11.7099	5.6954	1.4917
射雕英雄传	11.8224	5.485	1.272
白马啸西风	10.2789	3.9877	0.7365
碧血剑	11.7309	5.0051	1.0017
神雕侠侣	11.6801	5.4905	1.339
笑傲江湖	11.3827	5.6221	1.5396
越女剑	10.1031	2.5278	0.3243
连城诀	11.0263	4.7036	0.962
雪山飞狐	11.0988	4.1173	0.6991
飞狐外传	11.513	5.0008	1.0627
鸳鸯刀	10.4994	3.0937	0.4279
鹿鼎记	11.4413	5.7694	1.6213

四．总结

通过本次作业，对语言的处理有了初步的认识，了解并尝试完成了一个关于语言信息熵计算的小程序。在过程中遇到了诸如 txt 文字格式需要修改等困难，通过资料的查询等将其解决。本次作业也参考了部分 github 上的代码等，有了不小的收获。