Report of Deep Learning for Natural Langauge Processing

冯佳铂
775193638@qq.com

Part 1 ： 验证齐夫定律

Abstract

对金庸先生 16 篇小说建立中文语料库验证齐夫定律（Zip’s law），计算中文(分别以词和字为单位）的平均信息熵。

Introduction

齐夫定律（Zip’s law）是由哈佛大学的语言学家乔治·金斯利·齐夫（George Kingsley Zipf）于 1949 年发表的实验定律。它表述为：在自然语言的语料库里，一个单词出现的频率与它在频率表中的排名成反比。

Methodology

.

M1: 验证齐夫定律

利用python读取16部小说，将其内容存放于file_text中，替换其中明显的干扰项，本书来自www.cr173.com免费txt小说下载站、标点符号、换行符、空格、更多更新免费电子书请关注www.cr173.com等，使用jieba库函数进行分词，制作stopwords.txt去除文本中常见无意义词汇，提高处理效率，减少噪音干扰。然后统计词与词频绘图观察验证定律。

Experimental Studies

T1: 验证齐夫定律

Conclusions

近似满足以坐标轴为对数尺的齐夫分布曲线。

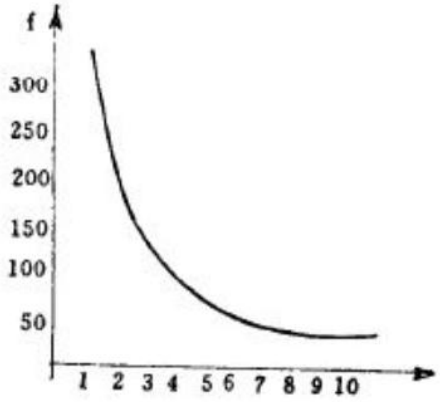


图 4—4 齐夫词频分布曲线

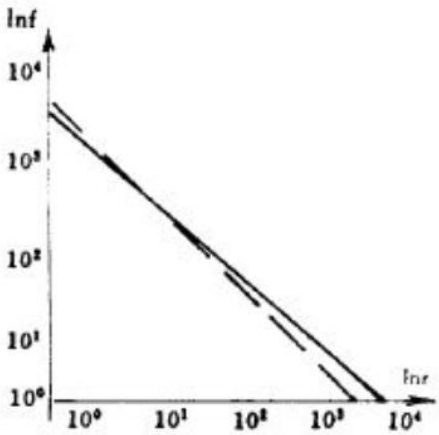


图 4—5 坐标轴为对数尺的齐夫分布曲线

References

[1] 刘鸿博.《Python 自然语言处理》.[Z].undefined.电子工业出版社.2022.
[2]https://blog.csdn.net/sinat_41858359/article/details/125130804?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-125130804-blog-116195510.235%5Ev43%5Epc_blog_bottom_relevance_base4&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-125130804-blog-116195510.235%5Ev43%5Epc_blog_bottom_relevance_base4&utm_relevant_index=11
[3]https://blog.csdn.net/dilifish/article/details/117885706

Part 2 ： 计算中文平均信息熵

Abstract

对金庸先生 16 篇小说建立中文语料库，并计算其中文平均信息熵。

Introduction

信息是个很抽象的概念。人们常常说信息很多，或者信息较少，但却很难说清楚信息到底有多少。直至香农提出了“信息熵”的概念，才解决了对信息的量化度量问题。分别利用基于词的一元模型、二元模型和三元模型计算所给中文语料库的信息熵。

Methodology

.

M1：一元模型

$H(x) = -\sum_{x \epsilon X} P(x) \log P(x)$，其中$P(x)$近似等于每个词在语料库中出现的频率。

M2：二元模型

$H(X|Y) = -\sum_{x \epsilon X, y \epsilon Y} P(x,y) \log P(x|y)$，其中$P(x,y)$近似等于每个二元词在语料库中出现的频率，条件概率$P(x|y)$近似等于每个二元词组在语料库中出现的频数与该二元词组的第一个词为词首的二元词组的频数的比值。

M3：三元模型

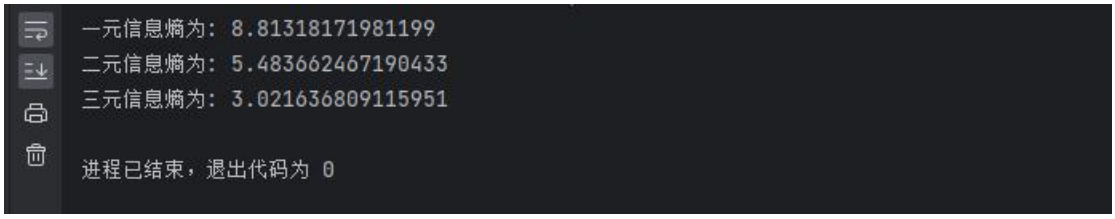
$H(X|Y,Z) = -\sum_{x \epsilon X, y \epsilon Y, z \epsilon Z} P(x,y,z) \log P(x|y,z)$，其中$P(x,y,z)$近似等于每个三元词组在语料库中出现的频率，条件概率$P(x|y,z)$近似等于每个三元词组在语料

库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

Experimental Studies

T1：验证齐夫定律

模型	词信息熵
一元模型	8.8132
二元模型	5.4837
三元模型	3.0216



Conclusions

随着模型复杂度的增加，信息熵下降，信息熵的值越大，文本中不同字符或词组的出现频率相差越大，文本的特征值越分散；模型变复杂后，考虑词与词之间的搭配，上下文之间的联系，能够更准确的扑捉到某些语言现象，从而使得这些现象的出现变得更加可预测，所以信息熵降低。

References

[1] 刘鸿博.《Python 自然语言处理》.[Z].undefined.电子工业出版社.2022.

[2]中文信息熵的计算_汉语信息熵-CSDN 博客[3]

[3]Brown, P. F. , Pietra, S. D. A. , Pietra, V. D. J. , Lai, J. C. , & Mercer, R. L. . (1992). An estimate of an upper bound for the. Computational Lingus, 18(1), 31-40.