深度学习与自然语言处理作业第一次作业

汪婧昀
19231136@buaa.edu.cn

实验任务

1.通过中文语料库来验证 Zipf's Law.
2.阅读 Entropy Of English，计算中文(分别以词和字为单位) 的平均信息熵。

实验原理

一、Zipf's Law

在给定的语料中，对于任意一个 term，其频度(freq)的排名（rank）和频度（freq）的乘积大致是一个常数。

二、信息熵

1、熵

熵，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。亦被社会科学用以借喻人类社会某些状态的程度。熵的概念是由德国物理学家克劳修斯于 1865 年所提出。最初是用来描述“能量退化”的物质状态参数之一，在热力学中有广泛的应用。但那时熵仅仅是一个可以通过热量改变来测定的物理量，其本质仍没有很好的解释，直到统计物理、信息论等一系列科学理论发展，熵的本质才逐渐被解释清楚，即，熵的本质是一个系统“内在的混乱程度”。

2、信息熵

信息熵的公式如下：

$$H(x) = -\sum p(x) \log p(x)$$

信息论之父克劳德·香农给出了信息熵的三个性质：

1）单调性，发生概率越高的事件，其携带的信息量越低；

2）非负性，信息熵可以看作为一种广度量，非负性是一种合理的必然；

3）累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

三、N-Gram 语言模型

1、语言模型

对于自然语言相关的问题，比如机器翻译，最重要的问题就是文本的序列有时候不是符合我们人类的使用习惯，语言模型就是用于评估文本序列符合人类语言使用习惯程度的模型。当前的语言模型是以统计学为基础的统计语言模型，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。

给定一个句子（词语序列）如下：

$$S = W_1, W_2, ..., W_k$$

其出现概率可以表示为：

$$P(S) = P(W_1, W_2, ..., W_k) = P(W_1)P(W_2|W_1)...P(W_k|W_1, W_2, ..., W_{k-1})$$

但是事实上不能用这种方式去计算条件概率，直接这样计算会导致参数空间过大和数据稀疏严重。假设一个语料库中单词的数量为|V|个，一个句子由 n 个词组成，那么每个单词都可能有|V|个取值，那么由这些词组成的 n 元组合的数目为|V|×n 种，也就是说，组合数会随着 n 的增加而呈现指数级别的增长，随着 n 的增加，预料数据库能够提供的数据是非常有限的，除非有海量的各种类型的语料数据，否则还有大量的 n 元组合都没有在语料库中出现过（即由 n 个单词所组成的一些句子根本就没出现过，可以理解为很多的 n 元组所组成的句子不能被人很好地理解）也就是说依据最大似然估计得到的概率将会是 0，模型可能仅仅能够计算寥寥几个句子。由此需要引入 N-Gram 语言模型。

2、N-Gram 语言模型

引入马尔科夫假设：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。将上面的计算方式是通过马尔科夫假设进行简化的，假设词语只与它前面的 k 个词语相关，就得到条件概率计算简化如下：

$$P(W_i|W_1, W_2, ..., W_{i-1}) = \prod_i P(W_i|W_{i-k}, ..., W_{i-1})$$

一元模型、二元模型、三元模型对应的 k 分别为 0，1，2，即 N-Gram 语言模型对应的 k 为 N-1。

实验内容

一、验证 Zipf's Law

使用 jieba 库对中文语料库进行分词，统计除停词外其他词的出现频度。

二、以词为单位计算平均信息熵

使用 jieba 库对中文语料库进行分词。

一元模型只需要统计每个词在语料库中出现的频数，得到词频表。

二元模型也需要统计每次词在语料库中出现的频数，得到词频表，作为计算条件概率 P(wi|wi-1)时的分母，并且需要统计每个二元词组在语料库中出现的频数，得到二元模型词频表。

三元模型需要统计每个二元词组在语料库中出现的频数，得到二元模型词频表，作为计算条件概率 P(wi|wi-2,wi-1)时的分母，并且需要统计每个三元词组在语料库中出现的频数，得到三元模型词频表。

三、以字为单位计算平均信息熵

将文本以字为单位切分，其余同上。

实验结果

一、验证 Zipf's Law

如图 1 所示，不考虑停词，在中文语料库中以词为单位统计，词出现频度的排名和其频度的乘积大致为一常数，表现在图上为成反比。

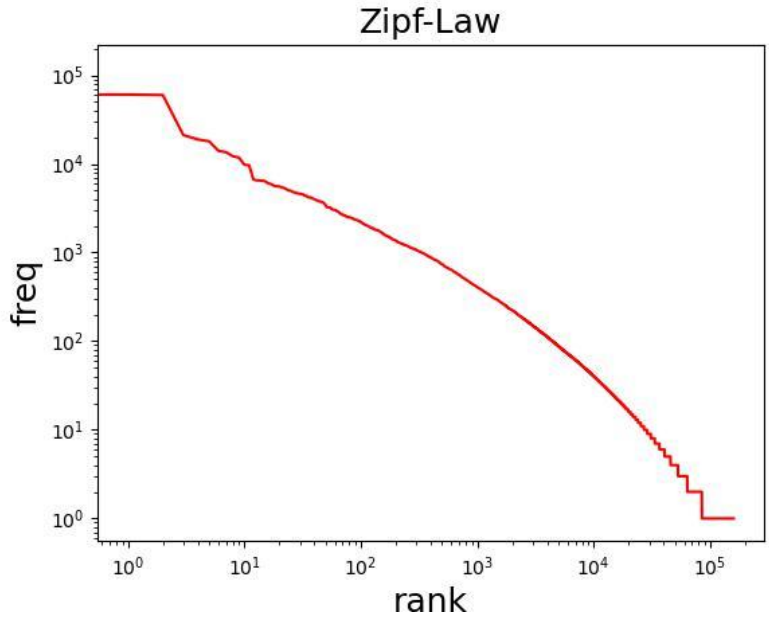

图 1：验证 Zipf's Law

二、以词/字为单位计算平均信息熵

对比 1-gram、2-gram、3-gram 三种语言模型得到的结果可以看到，N 取值越大，即考虑前后文关系的长度越大，不同词出现的个数越多，这是因为长度的增加也增加了由字组合成词的组合个数，所以会出现更多不同的词。而随着 N 取值变大，文本的信息熵则越小，这是因为 N 取值越大，通过分词后得到的文本中词组的分布就越简单，N 越大使得固定的词数量越多，固定的词能减少由字或者短词打乱文章的机会，使得文章变得更加有序，减少了由字组成词和组成句的不确定性，也即减少了文本的信息熵，符合实际认知。

以词为单位的实验结果如表 1 所示：

表 1: 以词为单位计算平均信息熵

语言模型	词库总词数	信息熵（比特/词）
一元模型	4314429	12.2
二元模型	4255111	6.9
三元模型	4196042	2.3

以字为单位的实验结果如表 2 所示：

表 2: 以字为单位计算平均信息熵

语言模型	词库总字数	信息熵（比特/字）
一元模型	7299812	9.5
二元模型	7240494	6.7
三元模型	7181176	3.9