北京邮电大学 计算机学院

《自然语言处理》实验报告

姓名 王睿嘉

学号2015211906

班级2015211307

N-gram Language Models

1. 实验内容和环境描述
2. 实验内容

**N-gram Language Models: 20 points**

-In this assignment you will explore a simple, typical N-gram language model.

-This model can be trained and tested on sentence-segmented data of a Chinese text corpus. “Word Perplexity” is the most widely-used evaluation metric for language models.

-Additional points: if you can test how does the different "Word Perplexity" of the different “N” grams, you will get additional 10 points

-Additional points: if you can test how does the different "Word Perplexity" of the different smoothing methods, you will get additional 10 points

1. 环境描述

编程语言：python

集成开发环境：PyCharm

解释器版本：2.7.13

1. 输入输出定义
2. 输入定义

用户无需自行输入数据。

本实验以压缩包所含文件“Chinese text corpus.txt”为语料库，通过一定划分产生训练样本和测试文本。

1. 输出定义

控制台分别输出trigram、bigram和trigram模型下，采用Add-one及Good-turing平滑方法所计算出的测试语料困惑度，与此同时，输出适当提示语句。

1. 实验设计
2. 整体说明

由于（词条，词频）的统计形式极为符合python字典数据结构，且re库中split函数便于语料库词条划分，因而采用python作为该实验的编程语言。

本实验使用教师所提供的“Chinese text corpus.txt”语料库，前90%的语料用于训练语言模型，剩余部分作为测试文本。

基本实现了unigram、bigram和trigram三种语言模型。由于存在数据稀疏问题（即受采样样本总数的限制，极大似然估计不可靠，或由于未知词和未知词组的存在，导致频率无法计算），对他们分别采用Add-one和Good-turing平滑方法，并计算出各自的困惑度。

1. 重点说明

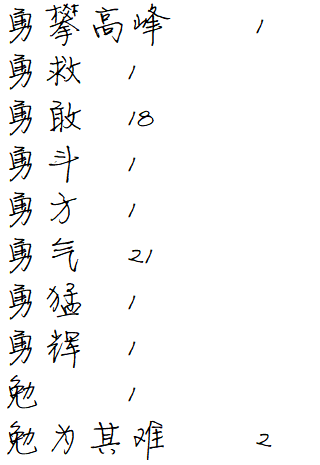
2.1 训练词典

在预处理过程中，对训练语料库进行词频统计，并将结果输出至unigram、bigram和trigram词典。

**Unigram词典格式：**

词 词频

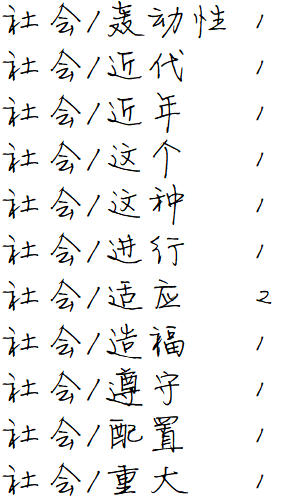
举例如下：



**Bigram词典格式：**

词/已出现词 词频

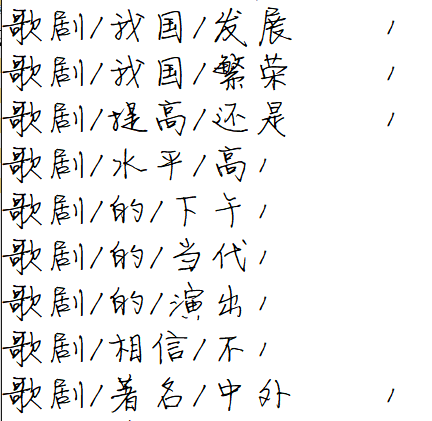
举例如下：



**Trigram词典格式：**

词/已出现词/更早出现词 词频

举例如下：

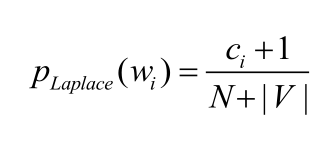


注：在输出bigram和trigram词典时，插入了句子开始标志“start”及结束标志“$”，并进行相应的词频统计。

2.2 Add-one平滑方法

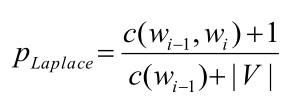
本实验对所有模型进行Add-one平滑处理。

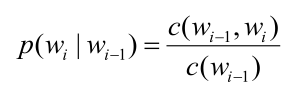
**Unigram计算公式：**



其中Ci为词Wi在训练预料中的出现次数，N为语料的token总数，|V|为unigram词表大小。

**Bigram计算公式：**



其中为已知词Wi-1出现的情况下词Wi出现的条件概率，|V|为bigram词表大小。

trigram计算方法与bigram同理。

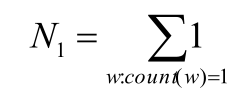
2.3 Good-turing平滑方法

本实验对所有模型进行Good-turing平滑处理。

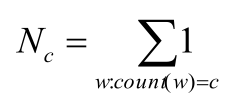
**Unigram计算方法：**

基本思路为根据仅出现一次的unigram个数，来确定那些未出现的unigram的概率。

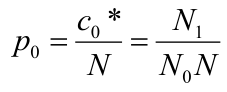
仅出现一次的unigram个数为



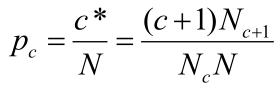
出现c次的unigram个数为



出现频次乘折扣可得，仅出现一次的unigram概率为

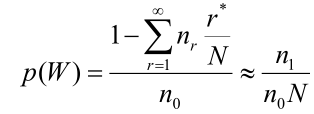


出现c次的unigram概率为

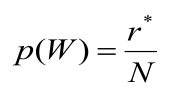
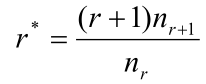


**扩展至N-gram中词W出现概率的计算方法：**

若其频次为0

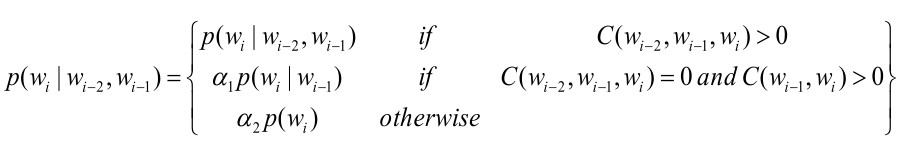


否则

其中，nr表示出现r次的N-gram个数。

注：在bigram与trigram的概率求解过程中，均采用回退，即高阶N-grams概率难以计算时，利用较低阶的N-grams来统计。以trigram的回退为例，套用以下公式：



其中，为了方便起见，bigram和unigram的概率计算均采用Add-one平滑方法。

|  |  |  |
| --- | --- | --- |
| 字段 | 值 | 含义 |
| OP | 02 | 该数据包为DHCP服务器送往客户端 |
| Htype | 01 | 硬件类型为以太网 |
| Hlen | 06 | 硬件的物理地址长度为6个字节，即以太网 |
| Hops | 00 | 客户端被设置为0，表示如果要用中继代理可以随意选择 |
| Transaction ID | 24075796 | 传输标识，一个由客户端选择的随机数，并在客户端和服务器通信中使用 |
| Seconds | 0000 | 客户端填写 |
| Ciaddr | 0000000 | 客户端IP地址，由于正在等待分配地址，所以该段用全0填写 |
| Yiaddr | 0a7af039 | 用来指定客户端的地址，DHCP分配的地址为10.122.240.57 |
| Siaddr | 0000000 | DHCP服务器的IP地址 |
| Giaddr | 0a7ac001 | 中继代理的IP地址，通过一个中继开机使用 |
| Chaddr | 08:d4:0c:58:11:6e | MAC地址 |
| Sname |  | 未给出 |
| File |  | 未给出 |

1. 实验结论

通过此次实验，进一步了解了网络中常用的各协议格式和工作原理，重点解析了IP、ICMP、DHCP、ARP、TCP等协议，并对其中一些关键字段和主要原理有了更深刻的理解和更理性的认识。

1. 实验心得

在本次实验中，遇到的主要问题有以下五点：

* 1. 根据实验指导书，先进行IP协议的抓包与分析。但在使用远程地址为[www.baidu.com](http://www.baidu.com的ping) 的ping命令时，所产生的IP数据包均传输在IPV6协议下。然而对IPV6协议仅为了解，无需细致掌握，便困惑如何才能捕获到IPV4数据包。带着疑问，先进行了DHCP协议的分析，在此过程中发现将数据发送至DHCP中继代理路由器，所产生的IP数据包为IPV4数据包，该问题得到解决；
  2. IP协议抓包结束后，擅自将结果以协议为关键字进行了排序，导致始终找不到最后一个分段数据包。随后，又按照正常的时间顺序查看结果，发现最后一个数据包为ICMP包，该问题得到解决；
  3. 验证IP分段原理时，对于最后一个ICMP包，遗忘了其封装在IP包中，还存在20字节的IP包头，导致各段数据相加非8000字节，与同学讨论后发现该问题，得到解决；
  4. TCP协议的抓包过程耗费较长时间。起初所选网址不当，导致未出现清晰的两端口数据传输，后更换网址，该问题得到解决；
  5. TCP实际的连接释放过程与课本上所讲解的有出入。书本上的释放过程大致同连接建立过程，经3次握手完成释放，然而，抓包结果通常是经4次握手，两个传输方向分别释放。

实践出真知，本次协议数据的捕获和解析实验是对课堂和书本所学知识的补充。网络中的实际情形与已了解到的原理大体一致，但又复杂许多。通过自己动手、亲力亲为捕获数据包，并对16进制数据进行分析，加深了对包头各字段功能的理解和记忆，收获颇丰。