lengyuhong的专栏

仰望天空，脚踏实地

个人资料

lengyuhong

访问： 971896次
积分： 11095
等级： BLOG 7
排名： 第1269名

原创： 175篇 转载： 160篇
译文： 2篇 评论： 106条

N-gram模型

标签： 语言 输入法 搜狗 游戏 工作 微软

2010-11-19 17:26 34663人阅读 评论(5)

分类： 搜索引擎（43）

N-Gram是大词汇连续语音识别中常用的一种语言模型，对中文而言，我们
Language Model)。汉语语言模型利用上下文中相邻词间的搭配信息，在需要
字母或笔划的数字，转换成汉字串(即句子)时，可以计算出具有最大概率的句
需用户手动选择，避开了许多汉字对应一个相同的拼音(或笔划串，或数字串)

该模型基于这样一种假设，第n个词的出现只与前面N-1个词相关，而与其
各个词出现概率的乘积。这些概率可以通过直接从语料中统计N个词同时出现
和三元的Tri-Gram。

文章分类

C/C++/C# (5)
google (3)
JAVA (82)
Linux (30)
PHP (0)
专业 收获 心得 (11)
其他 (3)
搜索引擎 (44)
数据库 (33)
牛人故事 (2)
算法 (6)

文章存档

2011年05月 (2)
2011年04月 (2)
2011年03月 (17)
2011年02月 (17)
2011年01月 (2)

展开

阅读排行

N-gram模型 (34663)
java对象的强引用，软引 (33909)
linux下用rpm 安装jdk

在介绍N-gram模型之前，让我们先来做个香农游戏（Shannon Game）。我们给定一个词，然后猜测下一个词是什么。当我说“艳照门”这个词时，你想到下一个词是什么呢？我想大家很有可能会想到“陈冠希”，基本上不会有人会想到“陈志杰”吧。N-gram模型的主要思想就是这样的。

对于一个句子T，我们怎么算它出现的概率呢？假设T是由词序列W1,W2,W3,…Wn组成的，那么P(T)=P(W1W2W3Wn)=P(W1)P(W2|W1)P(W3|W1W2)…P(Wn|W1W2…Wn-1)

补充知识：

3. 条件概率

$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

乘法公式

$$P(AB) = P(A)P(B \mid A) \quad (P(A) > 0)$$

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_1A_2 \cdots A_{n-1})$$
$$(P(A_1A_2 \cdots A_{n-1}) > 0)$$

但是这种方法存在两个致命的缺陷：一个缺陷是参数空间过大，不可能实用

为了解决这个问题，我们引入了马尔科夫假设：一个词的出现仅仅依赖于它

Java中怎样判断一个字符 (32637)
java虚拟机内存管理机制 (32600)
Log4J入门教程（一）入 (21249)
运用Jconsole监控JVM (19696)
Jboss调优——最佳线程 (19482)
linux vmstat命令详解 (15409)
java相对路径设置 (15288)
(14798)

评论排行

Log4J入门教程（一）入 (13)
N-gram模型 (5)
java对象的强引用，软引 (4)
Jboss调优——最佳线程 (4)
Linux脚本（shell）编程 (3)
Log4J入门教程（二）参 (3)
对称密码算法DES中的子 (3)
Java开源分词系统IKAna (3)
linux下用rpm 安装jdk (3)
linux top指令 (2)

推荐文章

* 程序员4月书讯：Angular来了！
* CSDN日报20170516 ——《驱动小白和硬件老司机关于硬件那点事儿的一次密谈》

如果一个词的出现仅依赖于它前面出现的一个词，那么我们就称之为bigram。即

P(T) = P(W1W2W3…Wn)=P(W1)P(W2|W1)P(W3|W1W2)…P(Wn|W1W2…Wn-1)

≈P(W1)P(W2|W1)P(W3|W2)…P(Wn|Wn-1)

如果一个词的出现仅依赖于它前面出现的两个词，那么我们就称之为trigram。

在实践中用的最多的就是bigram和trigram了，而且效果很不错。高于四元的用的很少，因为训练它需要更庞大的语料，而且数据稀疏严重，时间复杂度高，精度却提高的不多。

那么我们怎么得到P(Wn|W1W2…Wn-1)呢？一种简单的估计方法就是最大似然估计(Maximum Likelihood Estimate）了。即P(Wn|W1W2…Wn-1) = (C(W1 W2…Wn))/(C(W1 W2…Wn-1))

剩下的工作就是在训练语料库中数数儿了，即统计序列C(W1 W2…Wn) 出现的次数和C(W1 W2…Wn-1)出现的次数。

下面我们用bigram举个例子。假设语料库总词数为13,748

I	3437
want	1215
to	3256
eat	938
Chinese	213
food	1506
lunch	459

表 1 词和词频

* 简单粗暴地入门机器学习

* AntShares区块链的节点部署与搭建私有链

* 分布式机器学习的集群方案介绍之HPC实现

* Android 音频系统：从 AudioTrack 到 AudioFlinger

	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

表 2 词序列频度

P(I want to eat Chinese food)

=P(I)*P(want|I)*P(to|want)*P(eat|to)*P(Chinese|eat)*P(food|Chinese)

=0.25*1087/3437*786/1215*860/3256*19/938*120/213

=0.000154171

ps：网上很多资料中，表1，词与词频的张表是没有的，所以造成文章表意

这里还有一个问题要说，那就是数据稀疏问题了，假设词表中有20000个词有400000000个，如果是trigram，那么可能的N-gram就有8000000000000个！那料库中都没有出现，根据最大似然估计得到的概率将会是0，这会造成很大的项为0，那么整个句子的概率就会为0，最后的结果是，我们的模型只能算可怜得的概率是0. 因此，我们要进行数据平滑（data Smoothing），数据平滑的目的之和为1，使所有的N-gram概率都不为0.有关数据平滑的详细内容后面会再讲到

freezenfire: 非常感谢很好的文章！！这么好的文章只有谷歌能搜到。。。

藏经阁（关注的博客）

淘宝数据平台团队
民间推荐系统的组织者谷文栋博客
中科院自动化所项亮博士的推荐系统主题博客
百分点推荐技术研究中心
刘未鹏
淘宝Tair键值数据库
阿里巴巴数据库团队

归云庄（研究领域）

Java虚拟机（JVM）
中文分词
MySQL
java基础
myeclipse
shell
Linux
lucene
算法
数据库

乾坤大挪移

我的新浪微薄

了解了噪声信道模型和N-gram模型的思想之后，其实我们自己就能实现一个音词转换系统了，它是整句**智能**输入法的核心，其实我们不难猜到，搜狗拼音和微软拼音的主要思想就是N-gram模型的，不过在里面多加入了一些语言学规则而已。

顶 6　　踩 0

上一篇　【转】基于统计的词网格分词
下一篇　淘宝的新长征

相关文章推荐

- Deep-Learning paper整理
- threejs 源码注释二十五CoreGeometryjs
- Python 网页爬虫 & 文本处理 & 科学计算 & 机器学习...
- 28款GitHub最流行的开源机器学习项目
- 28款GitHub最流行的开源机器学习项目
- Python 开源项目
- 基于词表和N-gram
- 基于N-gram的双向
- InnoDB全文索引N

参考知识库

语音识别与合成知识库

671 关注 | 320 收录

人工智能规划与决策知识库

598 关注 | 8 收录

猜你在找

微软专家C语言系列之 文件细节讲解与数据检索

微软专家C语言系列之 文件处理

VR实战案例 | HTC Vive射箭游戏

Cocos2d-x 3.x 项目实战：仿微信飞机大战(射击类游戏)

教你玩转游戏制作Construct2

R语言在游戏数据行业的应用

使用Cocos开发一款简单的3D VR抓钱游戏

20150618.CPP语言

20150619.CPP语言

20150613.CPP语言

查看评论

您还没有登录,请[登录]或[注册]

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

核心技术类目