

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)



博客 (/blog.csdn.net/ref=toolbar) 学院 (/edu.csdn.net?ref=toolbar)

下载 (/download.csdn.net?ref=toolbar) GitChat (/gitbook.cn/?ref=csdn)

更多 ▾

weibo (/writeblogpostcontent?ref=toolbar)source=csdn

重新实现关于Mikolov的集成文本分类实验（详细过程） -

原创 2016年03月01日 23:01:53

标签：ubuntu (http://so.csdn.net/so/search/s.do?q=ubuntu&t=blog) /

linux (http://so.csdn.net/so/search/s.do?q=linux&t=blog) / NLP (http://so.csdn.net/so/search/s.do?q=NLP&t=blog)

759

前言：为了实现文本分类，将一个文本内容确定的分为积极或者消极，我们采用了Mikolov的文本分类方法，通过他在试验中的方法实现文本的二值分类。本文旨在如何重现他论文中实现的分类实验。论文参看Mikolov的ENSEMBLE OF GENERATIVE AND DISCRIMINATIVE TECHNIQUES FOR SENTIMENT ANALYSIS OF MOVIE REVIEWS

总的来说这次实验是磕磕绊绊，从配置环境，到改脚本命令，最后达到理想的实验结果。此文作为总结，将包含过程中使用到的方法、技术，以备后来使用。本文将分为四个部分来进行记录，将完成实验的总体步骤重新部署一下，整个实验进行下来预计在3到5个小时，安装系统，下载相关文件要1个多小时，训练数据分类数据要2个多小时左右。在虚拟机上进行实验可能需要8个小时或者更多。

一、实验准备

本文实验重现的文章是Mikolov的关于Classification的一篇Paper。Mikolov在文章中提供了他情绪分析-文本分类的实验代码，代码是在Linux环境下运行的，所以我们需要配置一个完整的适合该实验的环境。代码运用到了python，gcc环境，以及一些已有的算法分类器。由于他的代码是用到什么就下载什么，所以有的时候会出现下载错误的问题导致实验无法继续，所以我们需要事先下载好一些实验用到的算法包。以下将说明我们需要准备的一切工作：

- 1、Linux系统iso镜像（我使用的是Ubuntu14.04.3）
- 2、Liblinear-1.96.zip压缩包（我将提供在附件当中）百度即可下载
- 3、rnnlm-0.3e.tgz压缩包（我将提供在附件当中，此包经常会下载丢失，所以需要先下载好）百度即可
- 4、python numpy环境包，在http://sourceforge.net/projects/numpy/files中下载numpy包
- 5、虚拟机VMWare或者一台装了Linux的电脑（起初我用的是虚拟机，但是效率太低。所以最后借了二、Ubuntu下的环境配置

在虚拟机上安装系统或者装Linux系统我就不在此过多赘述，不是重点。

- 1、进入Ubuntu系统，调出终端Ctrl+Alt+T
- 2、输入sudo passwd回车会提示你输入你设置的系统密码，然后继续回车，此处自己设置新的root权限密码。
- 3、输入sudo apt-get install gcc/sudo apt-get install g++，下载并且安装gcc/g++环境。
- 4、将我们事先准备好的numpy包拷贝到系统中并且解压，通过cd命令进入解压后的路径，输入命令sudo python setup.py install安装此包（注意空格），在此之前需要键入sudo apt-get install python-dev安装此包。



种一颗牙多少钱 牙齿矫正价格表

望京soho 在职研究生取消 全口种植..  
种植牙的寿命 种牙多少钱一颗  
种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

广告 + 关注  
(http://blog.csdn.net/dyc773912355)  
码云

原创	粉丝	喜欢	未开通
7	0	0	(https://github.com/dyc773912355)

他的最新文章  
更多文章 (http://blog.csdn.net/dyc773912355)

Linux Shell命令（不定期更新）(http://blog.csdn.net/dyc773912355/article/details/50917990)

本地虚拟机Ubuntu14.04系统和宿主机Windows系统通信问题（java编写的socket通信）(http://blog.csdn.net/dyc773912355/article/details/50917160)

java Socket使用 (http://blog.csdn.net/dyc773912355/article/details/50908219)

配置Ubuntu14.04 下的java开发环境笔记 (http://blog.csdn.net/dyc773912355/article/details/50907965)

Stanford Segment 使用笔记 (http://blog.csdn.net/dyc773912355/article/details/50794212)

相关推荐

云摘录 | Word2Vec 作者Tomas Mikolov



种一颗牙多少钱 牙齿矫正价格表

望京soho 在职研究生取消 全口种植..  
种植牙的寿命 种牙多少钱一颗  
种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

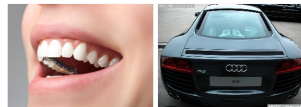
广告

立即体

验

内容举报

返回顶部



牙齿矫正价格表      奥迪r8二手

真火壁炉 隐形车衣 一颗牙齿多少钱

种植牙价格 种植牙寿命 种植牙的危害

植牙多少钱一颗 望京soho 做丰胸要..

OA办公系统 微型摄像机价格

## 达人课



望京soho 在职研究生取消 全口种植..  
种植牙的寿命 种牙多少钱一颗  
种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

Gy9YszLOzqMpgwBuV  
qqQhNP8vQiGIAPcmgIE  
种和为提供的中文词汇库进行文本分词过  
程中产生的中间文件log.csdn.net/dyc-  
tbojmg6dAnKcT0eRnI5r7090632)  
5HdZgnHNlmhkEusKzUjY  
kOAFVSHOOTZcqnoKDpyf

本地连接到数据库的IP地址和端口号  
46BPKWmLWysPcwDwCz  
hdows系统安装后默认打开的socket  
0agLUWYS0ZKA5HcsP6  
KWlThngnHnkHnc)

重新实现关于Mikolov的集成文本分类实验  
(详细过程) - (<http://blog.csdn.net/dyc73912355/article/details/50776390>)

## 配置Ubuntu14.04 下的java开发环境笔记

5、安装vim，输入sudo apt-get install vim

6、现在开始设置github的SSHkey，因为脚本中调用了github中的项目仓库中的代码。在终端中输入ssh-keygen

然后系统提示你保存SSH的位置，此时我们敲三次回车默认通过。然后系统会生成一个sshKey的文件保存在~/.ssh/id\_rsa.pub。此时我们键入命令 `vim ~/.ssh/id_rsa.pub` 打开文件，全选文件中的字符，从ssh-rsa开始到最后一个字符，复制到我们新建的.md文件中暂作保存。

接着拷贝`.ssh/id_rsa.pub`文件内的所有内容，将它粘贴到github帐号管理中的添加SSH key界面中。

打开github帐号管理中的添加SSH key界面的步骤如下：

1. 登录github
2. 点击右上方的Accounting settings图标
3. 选择 SSH key
4. 点击 Add SSH key

在出现的界面中填写SSH key的名称，填一个你自己喜欢的名称即可，然后将上面拷贝的~/.ssh/id\_rsa.pub文件内容粘贴到key一栏，在点击“add key”按钮就可以了。

添加过程github会提示你输入一次你的github密码。

添加完成后再次执行git clone就可以成功克隆github上的代码库了。

### 三、实验步骤

1、将Mikolov提供的iclr15文件拷贝到Ubuntu系统当中。将我们准备好的Liblinear和rnnlm包拷贝到Ubuntu当中以备使用。

2、接下来我们修改一些脚本代码，因为原始代码会删除一些下载好的包，我们需要这些包所以需要修改。

首先进入iclr15/scripts，找到data.sh脚本打开，将其中包含rm的移除命令代码全部删除或者用#注释掉。

进入install\_liblinear.sh文件，将wget一行代码删除或者注释掉，同时将rm命令代码删除或者注释掉。

进入rnnlm.sh文件，将wget一行代码删除或者注释掉，同时将rm命令代码删除或者注释掉。

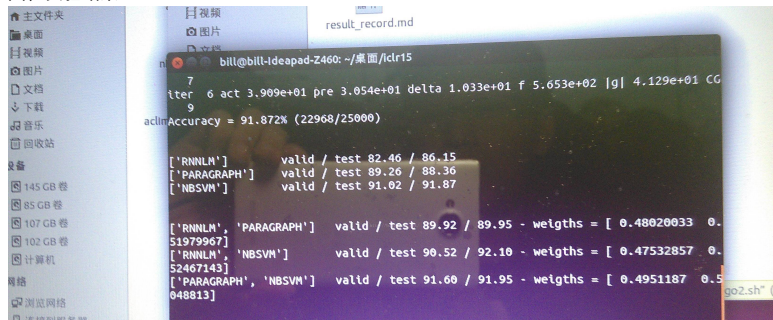
进入paragraph.sh文件，将rm命令代码删除或者注释掉。

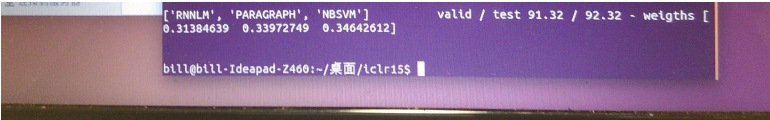
3、在iclr15文件夹所在目录建立一个新的文件夹命名为iclr15\_run，在其中建立rnnlm文件夹。然后将liblinear包拷贝到iclr15\_run目录下，将rnnlm包拷贝到rnnlm文件夹中。

4、调出终端，用cd命令进入到iclr15文件夹，键入chmod +x oh my go.sh。然后键

5、整个实验过程是，在斯坦福网站上下载训练数据（影评数据），将下载好的分类数据分配到train，test，unsup文件集中以备使用，调用nnml算法模型训练数据，并且测试数据。调用word2vec（作者基于此提出的paragraphVec分类方法）训练并测试数据。调用NB-SVM算法模型训练并测试数据，通过ensemble整合测试效果，得出各个集成情况的评分。同时将集成的权重显示出来。更细致的集成过程此处没有讨论。

#### 四、实验结果





原文链接: <http://blog.csdn.net/dyc773912355/article/details/50907965>  
116

五、关于训练好的分类器的使用

此处我用于测试的分类器使用的是Mikolov已经提出很久的rnnlm的分类器。当模型训练好之后，再次用此模型训练时在模型根目录下键入命令：`./rnnlm -train train -valid valid -rnnlm model -hidden 1` 测试文件时在模型根目录下键入命令：`./rnnlm -rnnlm model -test test -debug 0 -nbest > modelScore`

其中 -train：训练文件

-valid：校验集的名称（一般为训练文件中的一小部分）

-rnnlm：输出模型的名称

-hidden：隐含层神经元个数

-debug：控制开关，设置值不同会提供一些输出，设为2会输出运行时参数。

-bptt：控制通过环反向传播错误。

-class：指定单词的分类。100表示分为100类。

-test：测试文件。

-rand-seed：指定随机种子，用来初始化网络的权值的，比如指定为1,那么内部会执行srand(1)，网络

-direct-order：这个参数是指定rnn中me（最大熵模型）部分特征的阶数。最大是不会超过20的，超

-binary：这个参数如果没有，则默认为text方式，区别在于binary是用二进制方式存储数据，text是以ascii方式，对于大量的浮点数来说，binary能更省文件大小，但是缺点是用文件打开，里面都是乱码。

-direct：这个参数的含义就比较技术细节了，它来指定网络输入层到输出层所存放权值的一维数组的大小，并且单位是一百万，比如现在指定的值为2,其内部大小是2000000。



望京soho 在职研究生取消 全口种植..  
种植牙的寿命 种牙多少钱一颗  
种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

广告


 发表你的评论

([http://my.csdn.net/weixin\\_35068028](http://my.csdn.net/weixin_35068028))

相关文章推荐


云摘录 | Word2Vec 作者Tomas Mikolov 的三篇代表作解析 ([http://blog.csdn.net/sinat\\_2691...](http://blog.csdn.net/sinat_2691...))

本文来源于公众号paperweekly 谈到了word2vec作者的三篇论文： 1、Efficient Estimation of Word Representation...

 [sinat\\_26917383](http://blog.csdn.net/sinat_26917383) ([http://blog.csdn.net/sinat\\_26917383](http://blog.csdn.net/sinat_26917383)) 2016年09月18日 20:39 3566

Scikit-learn实战之最近邻算法 (<http://blog.csdn.net/u013709270/article/details/53819741>)

1. 最近邻的概念 sklearn.neighbors 提供了基于最近邻的无监督和有监督学习方法的功能。无监督最近邻是许多其他学习方法的基础，尤其是流型学习和谱聚类。有监督的最近邻学习有两...

 [u013709270](http://blog.csdn.net/u013709270) (<http://blog.csdn.net/u013709270>) 2016年12月22日 20:36 787



2017年前端报告：程序员薪酬上涨70%！

 内容举报

 返回顶部

 内容举报

 返回顶部



前端程序员的薪酬曝光，2017年，平均上涨70%，月薪20的人最为常见！以下为详细数据....

([http://www.baidu.com/cb.php?c=lgF\\_pyfqhHmknj0dP1f0IZ0qnfK9ujYzP1nYPH0k0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YkrHb3rHmYuWN9nWDLmH9B0AwY5HDdnHnYnjD1rHT0lgF\\_5y9YIZ0IQzq-uZR8mLPbUB48ugfElAqspynElvNBnHqdlAdxTvqdThP-5yF\\_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqhHRLPjnvnfKEpyfqhHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP](http://www.baidu.com/cb.php?c=lgF_pyfqhHmknj0dP1f0IZ0qnfK9ujYzP1nYPH0k0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YkrHb3rHmYuWN9nWDLmH9B0AwY5HDdnHnYnjD1rHT0lgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfElAqspynElvNBnHqdlAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqhHRLPjnvnfKEpyfqhHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP))

文本挖掘之文本聚类的介绍以及应用 (<http://blog.csdn.net/u011955252/article/details/50803...>)

文本聚类是一个将文本集分组的全自动处理过程，是一种典型的无指导的机器学习过程。类是通过相关数据发现的一些组，类内的文本和其它组相比更为相近。换一种说法就是，文本聚类的目标是找到这样一些类的集合，类之间...

u011955252 (<http://blog.csdn.net/u011955252>) 2016年03月04日 15:53 955

文本分类实验中用java实现取名词和去除停用词 (<http://blog.csdn.net/yoany/article/details/4...>)

这小程序我调了一整天啊啊啊啊 本来打算用C语言写一个 可是弄了半天总是出错 上午几乎就在用C语言写了...

yoany (<http://blog.csdn.net/yoany>) 2014年11月17日 20:47 2770

社会化搜索与推荐浅析-朴素贝叶斯+laplace平滑文本分类器推导过程及java版实现 (<http://blog...>)

本文由larrylgq编写，转载请注明出处：<http://blog.csdn.net/larrylgq/article/details/7395261> 作者:吕桂强 邮箱: [larry.lv.wor...](mailto:larry.lv.wor...)

larrylgq (<http://blog.csdn.net/larrylgq>) 2012年03月26日 18:28 7156



一学就会的 WordPress 实战课

学习完本课程可以掌握基本的 WordPress 的开发能力，后续可以根据需要开发适合自己的主题、插件，打造最个性的 WordPress 站点。

([http://www.baidu.com/cb.php?c=lgF\\_pyfqhHmknjfvP1m0IZ0qnfK9ujYzP1f4Pjnz0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YknWP-mvR3nAP9mhFbmWN-0AwY5HDdnHnYnjD1rH60lgF\\_5y9YIZ0IQzqMpgwBUvqoQhP8QvGIAPCmgfEmvq\\_lyd8Q1N9nHmvnj7hnHPWnjFhPAD1Pyn4uW99ujqdlAdxTvqdThP-5HDknWw9mhhEusKzujYk0AFV5H00TZcqn0KdpyfqhHRLPjnvnfKEpyfqhHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPHc1rHf](http://www.baidu.com/cb.php?c=lgF_pyfqhHmknjfvP1m0IZ0qnfK9ujYzP1f4Pjnz0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YknWP-mvR3nAP9mhFbmWN-0AwY5HDdnHnYnjD1rH60lgF_5y9YIZ0IQzqMpgwBUvqoQhP8QvGIAPCmgfEmvq_lyd8Q1N9nHmvnj7hnHPWnjFhPAD1Pyn4uW99ujqdlAdxTvqdThP-5HDknWw9mhhEusKzujYk0AFV5H00TZcqn0KdpyfqhHRLPjnvnfKEpyfqhHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPHc1rHf))

文本分类全过程实现 ([http://blog.csdn.net/qq\\_15706283/article/details/73530717](http://blog.csdn.net/qq_15706283/article/details/73530717))

最近在做文本分类方面的内容，之前接触数据挖掘的算法比较多一点，对自然语言处理领域基本上没有接触过。在做这一部分的内容的时候也是花了一些精力。用了一周的时间，将整个过程实现了一遍。我还是属于这个领域的菜...

qq\_15706283 ([http://blog.csdn.net/qq\\_15706283](http://blog.csdn.net/qq_15706283)) 2017年06月21日 12:00 60

文本分类程序的实现过程（C++语言）——特征选择的预处理 (<http://blog.csdn.net/monsion/a...>)

这几天在看一个文本分类的程序，写一下具体的实现过程。有的时候看了算法，感觉很明白了，但是自己实现的时候却又无从下手。这次从一个实际程序出发，或许能够更好的理解。 首先是训练数据集和测试数据集...

monsion (<http://blog.csdn.net/monsion>) 2012年09月15日 20:35 2448



基于质心的文本分类的分析与实验结果 (<http://download.csdn.net/download...>)



种一颗牙多少钱 牙齿矫正价格表

望京soho 在职研究生取消 全口种植..

种植牙的寿命 种牙多少钱一颗

种植牙的危害 OA办公系统 it培训机..

真火壁炉 修复双眼皮 切开双眼皮

广告

内容举报

返回顶部



种一颗牙多少钱 牙齿矫正价格表

望京soho 在职研究生取消 全口种植..

种植牙的寿命 种牙多少钱一颗





[\(http://download.csdn.net/download/palydawn/4309347\)](#)

2014年06月13日 06:50 370KB

下载

[中文文本分类实验 \(http://download.csdn.net/download/palydawn/4309347\)](#)

[\(http://download.csdn.net/download/palydawn/4309347\)](#)

2012年05月17日 16:13 383KB

下载

### 基于朴素贝叶斯分类器的文本分类算法的实现过程分析 (http://blog.csdn.net/XBWer/article/d...

基于朴素贝叶斯分类器的文本聚类算法（上）http://www.cnblogs.com/phinecos/archive/2008/10/21/1315948.html 基于朴素贝叶斯分类器的文...

XBWer (http://blog.csdn.net/XBWer) 2014年07月11日 23:41 1646

[贝叶斯算法实现文本分类器 \(http://download.csdn.net/download/suyuan66..](#)

[\(http://download.csdn.net/download/suyuan66..](#)

2011年12月16日 17:42 727KB

下载

[基于Bayes的新sgroup 18828文本分类器的Python实现 \(http://download....](#)

[\(http://download.csdn.net/download/suyuan66..](#)

2017年03月30日 10:41 129KB

下载

### CNN和RNN在文本分类过程中的区别整理 (http://blog.csdn.net/baoyan2015/article/details/6...

在最左边的输出层有两个channel，每个channel是一个二维的矩阵，矩阵的列的长度等于语句sentence的长度（也就是sente nce中的单词个数，通过padding使得待分类的每个sente...

baoyan2015 (http://blog.csdn.net/baoyan2015) 2017年03月21日 10:47 3191

[朴素贝叶斯算法文本分类JAVA实现 \(http://download.csdn.net/download/si...](#)

[\(http://download.csdn.net/download/si...](#)

2017年04月25日 18:57 1.59MB

下载

### scikit-learn：构建文本分类的“pipeline”简化分类过程、网格搜索调参 (http://blog.csdn.net/m...

前两篇分别将“加载数据”和“提取tf、tf-idf，进而构建分类器”，其实这个过程，vectorizer => transformer => classifier，早已被“scikit-learn p...

mmc2015 (http://blog.csdn.net/mmc2015) 2015年07月12日 21:21 1876

### 基于Bayes和KNN的新sgroup 18828文本分类器的Python实现 (http://blog.csdn.net/liujian...

基于Bayes和KNN的新sgroup 18828文本分类器的Python实现 向@yangliuy大牛学习NLP，这篇博客是数据挖掘-基于贝叶斯算法及KNN算法的新sgroup18...

liujiandu101 (http://blog.csdn.net/liujiandu101) 2016年06月21日 17:45 629

### 朴素贝叶斯的概率理论及其python代码实现文本分类的实例 (http://blog.csdn.net/gentelyang...

一：朴素贝叶斯是一种基于概率分布进行分类的方法，概率论是朴素贝叶斯的基础，之所以被称为朴素，而不是贝叶斯就是因为它在贝叶斯的基础上，增添了两个条件，一个是各特征之间相互独立，第二是每个特征同等重要。朴...

种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

广告

内容举报

返回顶部

望京soho 在职研究生取消 全口种植..  
种植牙的寿命 种牙多少钱一颗  
种植牙的危害 OA办公系统 it培训机..  
真火壁炉 修复双眼皮 切开双眼皮

广告

http://blog.csdn.net/dyc773912355/article/details/50776390



5/6



 gentelyang (<http://blog.csdn.net/gentelyang>) 2017年07月16日 09:14  219



文本分类的python实现-基于SVM算法 ([http://blog.csdn.net/wangyajie\\_11/article/details/62...](http://blog.csdn.net/wangyajie_11/article/details/62...))

描述 训练集为评论文本，标签为 pos,neu,neg三种分类，train.csv的第一列为文本content，第二列为label。可以单独使用SV C训练然后预测，也可以使用管道pipeline...

 wangyajie\_11 ([http://blog.csdn.net/wangyajie\\_11](http://blog.csdn.net/wangyajie_11)) 2017年03月15日 15:25  415

Tensorflow实现基于LSTM的文本分类方法 (<http://blog.csdn.net/feng98ren/article/details/78...>)

转载： <http://blog.csdn.net/u010223750/article/details/53334313?locationNum=7&fps=1> 引言 学习一段时...

 feng98ren (<http://blog.csdn.net/feng98ren>) 2017年11月19日 22:19  122

基于NaiveBayes的文本分类之Spark实现 ([http://blog.csdn.net/a\\_step\\_further/article/detail...](http://blog.csdn.net/a_step_further/article/detail...))

在尝试了python下面用sklearn进行文本分类（ [http://blog.csdn.net/a\\_step\\_further/article/details/50189727](http://blog.csdn.net/a_step_further/article/details/50189727) ）后，我们再来看下用spa...

 a\_step\_further ([http://blog.csdn.net/a\\_step\\_further](http://blog.csdn.net/a_step_further)) 2016年04月18日 08:08  1398



内容举报



返回顶部



0



种一颗牙多少钱

牙齿矫正价格表

望京soho 在职研究生取消 全口种植..

种植牙的寿命 种牙多少钱一颗

种植牙的危害 OA办公系统 it培训机..

真火壁炉 修复双眼皮 切开双眼皮



44 45 46 47