Python实现基于朴素贝叶斯的垃圾邮件分类 标签： python朴素贝叶斯垃圾邮件分类 04-20 15:09 2750人阅读 评论(1) 收藏 举报 分类： 机器学习（19） 听说

转载 starzhou 最后发布于2017-08-07 09:10:27 阅读数 605 ☆收藏

Python实现基于朴素贝叶斯的垃圾邮件分类

标签： python 朴素贝叶斯 垃圾邮件分类 2016-04-20 15:09 2750人阅读 评论(1

分类： 机器学习（19）

听说朴素贝叶斯在垃圾邮件分类的应用中效果很好，寻思朴素贝叶斯容易实现，就用**Python**写了一个朴素贝叶斯模型下的垃圾邮件分类。在400封邮件与垃圾邮件各一半）的**测试**集中测试结果为分类准确率95.15%，在仅仅统计词频计算概率的情况下，分类结果还是相当不错的。

实现代码及数据集下载

1、准备工作
python3.4开发环境；
结巴分词工具：https://github.com/fxsjy/jieba

2、贝叶斯公式
我们要做的是计算在已知词向量$w=(w_1,w_2,\dots,w_n)$ 的条件下求包含该词向量邮件是否为垃圾邮件的概率，即求：

$$P(s|w),w=(w_1,w_2,\dots,w_n)$$

其中，s表示分类为垃圾邮件
根据贝叶斯公式和全概率公式，

$$P(s|w_1,w_2,\dots,w_n)$$
$$=\frac{P(s,w_1,w_2,\dots,w_n)}{P(w_1,w_2,\dots,w_n)}$$
$$=\frac{P(w_1,w_2,\dots,w_n|s)P(s)}{P(w_1,w_2,\dots,w_n)}$$
$$=\frac{P(w_1,w_2,\dots,w_n|s)P(s)}{P(w_1,w_2,\dots,w_n|s)\cdot p(s)+P(w_1,w_2,\dots,w_n|s')\cdot p(s')} \quad \dots 式1$$

根据朴素贝叶斯的条件独立假设，并设先验概率$P(s)=P(s')=0.5$,上式可化为：

$$=\frac{\prod_{j=1}^{n}P(w_j|s)}{\prod_{j=1}^{n}P(w_j|s)+\prod_{j=1}^{n}P(w_j|s')}$$

再利用贝叶斯$P(w_j|s)=\frac{P(s|w_j)\cdot P(w_j)}{P(s)}$,式子化为

$$=\frac{\prod_{j=1}^{n}P(s|w_j)}{\prod_{j=1}^{n}P(s|w_j)+\prod_{j=1}^{n}P(s'|w_j)}$$
$$=\frac{\prod_{j=1}^{n}P(s|w_j)}{\prod_{j=1}^{n}P(s|w_j)+\prod_{j=1}^{n}(1-P(s|w_j))} \quad \dots 式2$$

至此，我们接下来会用式2来计算概率$P(s|w)$ ，为什么不用式1而用式2来计算概率，是因为通过式2可以将关于s'的部分用s表示，方便计算。

3、实现步骤
具体实现的源码已经给出，这里简单说下思路，就是一个分词并记录词频的过程：
（1）对训练集用结巴分词，并用停用表进行简单过滤，然后使用正则表达式过滤掉邮件中的非中文字符；
（2）分别保存正常邮件与垃圾邮件中出现的词有多少邮件出现该词，得到两个词典。例如词”疯狂”在8000封正常邮件中出现了20次，在8000封垃圾了200次；
（3）对测试集中的每一封邮件做同样的处理，并计算得到$P(s|w)$ 最高的15个词，在计算过程中，若该词只出现在垃圾邮件的词典中，则令$P(w|s')=$之亦然；若都未出现，则令$P(s|w)=0.4$ 。PS.这里做的几个假设基于前人做的一些研究工作得出的。
（4）对得到的每封邮件中重要的15个词利用式2计算概率，若概率> 阈值α(一般设为0.9) ，则判为垃圾邮件，否则判为正常邮件。

机器学习：朴素贝叶斯邮件分类（python实现） 阅读数 47
之前人工智能课程需要交一个小作业，参考网上文章，用贝叶斯做了一个邮件分类器，分享给大家。代码from re im... 博文 来自：Warms的博客

python_NLP实战之中文垃圾邮件分类 阅读数 2368
一、机器学习训练的要素数据、转换数据的模型、衡量模型好坏的损失函数、调整模型权重以便最小化损失函数的算... 博文 来自：t_zht的博客

python朴素贝叶斯电子邮件分类实例 阅读数 686
文章目录一、步骤二、实例+注释一、步骤（1）收集数据：提供文本文件。（2）准备数据：将文本文件解析成词条... 博文 来自：duanbin的博客

Python实现基于朴素贝叶斯的垃圾邮件分类 阅读数 1万+
听说朴素贝叶斯在垃圾邮件分类的应用中效果很好，寻思朴素贝叶斯容易实现，就用python写了一个朴素贝叶斯模型... 博文 来自：Kobe Bryant的专栏

朴素贝叶斯算法——实现垃圾邮件过滤（Python3实现） 阅读数 1万+
目录1、朴素贝叶斯实现垃圾邮件分类的步骤2、数据集下载3、代码实现4、朴素贝叶斯的优点和缺点1、朴素贝叶斯... 博文 来自：Asia-Lee的博客

基于朴素贝叶斯的垃圾分类算法（Python实现） 阅读数 3224
一、模型方法本工程采用的模型方法为朴素贝叶斯分类算法，它的核心算法思想基于概率论。我们称之为“朴素”，... 博文 来自：qq_39559641的博客

基于朴素贝叶斯+Python实现垃圾邮件分类和结果分析 阅读数 1856
基于朴素贝叶斯+Python实现垃圾邮件分类朴素贝叶斯原理请参考：贝叶斯推断及其互联网应用（二）：过滤垃圾邮... 博文 来自：Galoa的博客

基于朴素贝叶斯的垃圾邮件分类器Java实现和讲解 阅读数 930
朴素贝叶斯算法最典型的应用就是垃圾邮件的识别，在数据量非常大的情况下，识别的正确率可以达到接近100%，... 博文 来自：aGreySky的博客

python基于朴素贝叶斯算法实现新闻分类 阅读数 464
python基于朴素贝叶斯分类算法实现新闻分类 博文 来自：余夏婷的博客

机器学习算法三——基于概率论的分类方法：朴素贝叶斯（2）（示例：使用朴素贝叶斯过滤垃圾邮件） 阅读数 159
示例：使用朴素贝叶斯过滤垃圾邮件首先，将文本解析成词条；然后，和前面的分类代码集成为一个函数，该函数在... 博文 来自：Blog

【python与机器学习入门3】朴素贝叶斯2——垃圾邮件分类 阅读数 604
参考博客：朴素贝叶斯基础篇之言论过滤器 （po主Jack-Cui,《——大部分内容转载自 参考书籍：《机器... 博文 来自：momottyy的专栏

手把手教你用Python+朴素贝叶斯实现垃圾邮件分类 阅读数 2796
用朴素贝叶斯进行简单的垃圾邮件分类import numpy as npimport reimport osimport randomimport numpy as... 博文 来自：akira

02-29 朴素贝叶斯(垃圾邮件分类) 阅读数 26
文章目录朴素贝叶斯(垃圾邮件分类)邮箱训练集下载地址模块导入文本预处理遍历邮件训练模型测试模型朴素贝叶斯(... 博文 来自：小猿取经

python语言实现基于朴素贝叶斯算法的垃圾邮件过滤器 阅读数 476
引言应用python语言开发，采用交叉验证法，以收集的一些英文邮件作为语料，应用朴素贝叶斯分类方法。设先验概... 博文 来自：WangXin Progra...

python实现朴素贝叶斯算法 阅读数 726

python实现朴素贝叶斯算法朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。对于很多场景，朴素贝... 博文 来自：Wprofessor的博客

【机器学习基础】朴素贝叶斯进行垃圾邮件分类 阅读数 89

目录 一 朴素贝叶斯简介 二 贝叶斯决策理论 三 朴素贝叶斯进行垃圾邮件分类 3.1构造数据集 3.2 构造词典 3.3 构造... 博文 来自：Tuzi_bo的专栏

《机器学习实战》学习笔记：基于朴素贝叶斯的垃圾邮件过滤 阅读数 1万+

概率是许多机器学习算法的基础，在前面生成决策树的过程中使用了一小部分关于概率的知识，即统计特征在数据集... 博文 来自：liyuefeilong的专栏

基于朴素贝叶斯到中文垃圾邮件分类器 阅读数 6357

简介： 朴素 贝叶斯垃圾邮件分类器是在对邮件关键字进行统计分析到基础上利用贝叶斯公式进行分类到方法。相比... 博文 来自：Searching_Bird的...

基于朴素贝叶斯的文本分类算法 阅读数 421

基于朴素贝叶斯的文本分类算法 摘要：常用的文本分类方法有支持向量机、K-近邻算法和朴素贝叶斯。其中朴素贝叶... 博文 来自：kexinxin1的博客

基于朴素贝叶斯的垃圾邮件识别 阅读数 7859

在网上看到很多用朴素贝叶斯算法来实现垃圾邮件分类的，有直接调用库的，也有自己写的。出于对贝叶斯算法的复... 博文 来自：dayslrk的博客

朴素贝叶斯&基于朴素贝叶斯的文本分类算法 阅读数 487

朴素贝叶斯以及基于朴素贝叶斯的文本分类算法参考文章：https://www.cnblogs.com/jorbin/articles/1915888.ht... 博文 来自：G1011的博客

Python3 实现朴素贝叶斯分类 阅读数 275

Python3 实现朴素贝叶斯分类贝叶斯定理朴素贝叶斯代码分析测试贝叶斯定理贝叶斯定理是由已知事件概率和条件概... 博文 来自：Bcdfxg的博客

机器学习及python实现——朴素贝叶斯分类器 阅读数 2632

问题引入考虑构建一个垃圾邮件分类器，通过给定的垃圾邮件和非垃圾邮件的数据集，通过机器学习构建一个预测一... 博文 来自：Linkin_ygw的博客

在中国程序员是青春饭吗? 阅读数 8万+

今年，我也32了 ，为了不给大家误导，咨询了猎头、圈内好友，以及年过35岁的几位老程序员……舍了老脸去揭人家... 博文 来自：启舰

《MySQL 性能优化》之理解 MySQL 体系结构 阅读数 2万+

本文介绍 MySQL 的体系结构，包括物理结构、逻辑结构以及插件式存储引擎。 博文 来自：Tony.Dong的专栏

程序员请照顾好自己，周末病魔差点一套带走我。 阅读数 5万+

程序员在一个周末的时间，得了重病，差点当场去世，还好及时挽救回来了。... 博文 来自：敖丙

Python+OpenCV实时图像处理 阅读数 7万+

目录1、导入库文件2、设计GUI3、调用摄像头4、实时图像处理4.1、阈值二值化4.2、边缘检测4.3、轮廓检测4.4、... 博文 来自：不脱发的程序猿

2020年一线城市程序员工资大调查 阅读数 7万+

人才需求一线城市共发布岗位38115个，招聘120827人。其中beijing 22805guangzhou 25081shanghai 39614sh... 博文 来自：juwikuang的专栏

为什么猝死的都是程序员，基本上不见产品经理猝死呢? 阅读数 7万+

相信大家时不时听到程序员猝死的消息，但是基本上听不到产品经理猝死的消息，这是为什么呢？我们先百度搜一下... 博文 来自：曹银飞的专栏

害怕面试被问HashMap？这一篇就搞定了! 阅读数 4万+

声明：本文以jdk1.8为主！搞定HashMap作为一个Java从业者，面试的时候肯定会被问到过HashMap，因为对于Ha... 博文 来自：编码之外的技术博客

毕业5年，我问遍了身边的大佬，总结了他们的学习方法 阅读数 12万+

我问了身边10个大佬，总结了他们的学习方法，原来成功都是有迹可循的。 博文 来自：敖丙

python爬取百部电影数据，我分析出了一个残酷的真相 阅读数 4万+

2019年就这么匆匆过去了，就在前几天国家电影局发布了2019年中国电影市场数据，数据显示去年总票房为642.66... 博文 来自：Leo的博客

推荐10个堪称神器的学习网站 阅读数 23万+

每天都会收到很多读者的私信，问我：“二哥，有什么推荐的学习网站吗？最近很浮躁，手头的一些网站都看烦了，... 博文 来自：沉默王二

这些软件太强了，Windows必装！尤其程序员！ 阅读数 24万+
Windows可谓是大多数人的生产力工具，集娱乐办公于一体，虽然在程序员这个群体中都说苹果是信仰，但是大部... 博文 来自：编码之外的技术博客

阿里面试，面试官没想到一个ArrayList，我都能跟他扯半小时 阅读数 5万+
我是真的没想到，面试官会这样问我ArrayList。 博文 来自：敖丙

曾经优秀的人，怎么就突然不优秀了。 阅读数 5万+
职场上有很多辛酸事，很多合伙人出局的故事，很多技术骨干被裁员的故事。说来模板都类似，曾经是名校毕业，曾... 博文 来自：caoz的梦呓

C语言荣获2019年度最佳编程语言 阅读数 3010
关注、星标公众号，不错过精彩内容作者：黄工公众号：strongerHuang近日，TIOBE官方发布了2020年1月编程语... 博文 来自：strongerHuang

大学四年因为知道了这32个网站，我成了别人眼中的大神！ 阅读数 30万+
依稀记得，毕业那天，我们导员发给我毕业证的时候对我说“你可是咱们系的风云人物啊”，哎呀，别提当时多开心... 博文 来自：编码之外的技术博客

良心推荐，我珍藏的一些Chrome插件 阅读数 6万+
上次搬家的时候，发了一个朋友圈，附带的照片中不小心暴露了自己的 Chrome 浏览器插件之多，于是就有小伙伴评... 博文 来自：不忘初心

看完这篇HTTP，跟面试官扯皮就没问题了 阅读数 7万+
我是一名程序员，我的主要编程语言是 Java，我更是一名 Web 开发人员，所以我必须要了解 HTTP，所以本篇文章... 博文 来自：c旋儿的博客

python json java mysql pycharm android linux json格式

starzhou
TA的个人主页 >

原创	粉丝	获赞	评论	访问
401	857	502	171	196万+

等级: 博客 7 周排名: 965
积分: 2万+ 总排名: 532
勋章:

关注 私信

最新文章

bubbliiiing/keras-face-recognition
自然语言模型算法太杂乱？国产统一 AI 开源框架来了！
NUXT 入门第一课： 关于 Nuxt.js
前端渲染框架NUXT + UI组件 vertify
avue的使用

分类专栏

分析 111篇

举报

685篇

归档

热门文章

最新评论