

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

CSDN

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

0



登录 (https://passport.csdn.net/account/login?ref=toolbar) 注册 (https://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobile_register&source=csdnblog1)

standford自然语言处理第二课“文本处理基础 (Basic Text Processing) ”

翻译

2016年09月28日 21:42:55

161

1.机器学习中的单词positive和negative是什么含义？

positive是正例的意思，真真正正错的那个，例如：乳腺肿块；negative是错误的的意思，例如：正常。因为，我们主要是分出是否为乳腺肿块，所以肿块为正例。

2.NLP中的特征，特征粒度1.字符；2.词；3.人工特征

①字符：Lecun的深度学习实现文本分类，画网格，然后每个字母每个字母填充

②词

③人工特征：句号前后的单词是否大写开头、是否为缩略词、前后是否存在数字、句号前的单词长度、句号前后的单词在语料库中作为句子边界的概率等等。应用在断句中：

1.i am Dr.xu. %是一个句子label1

2.i am Dr.xu %不是一个句子label2

断句可以转化为分类问题，用好多分类器，但取决于多个人工好的特征。

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

立即体验



移民澳大利亚



二手豪华车



hahajing369 (http://blog.csdn.net/jingtingxu369)

+ 关注

(http://blog.csdn.net/jingtingxu369)

码云

原创

63

粉丝

5

喜欢

0

未开通

(https://github.com/jingtingxu369)

他的最新文章

更多文章 (http://blog.csdn.net/jingtingxu369)

matlab简单使用 (http://blog.csdn.net/jingtingxu369/article/details/78735203)

完形填空 (http://blog.csdn.net/jingtingxu369/article/details/78188056)

完形填空 (http://blog.csdn.net/jingtingxu369/article/details/78187504)

考研英语 (http://blog.csdn.net/jingtingxu369/article/details/76678291)



内容举报



返回顶部



3.未登录词识别：邱超的博物馆馆藏名，转化未分类问题。提取的人工特征：某个词x的前后是否为动词
观察发现：兵马俑出土于西安，收藏在西安博物馆。%“西安博物馆”为馆藏名，前面为“收藏”

二、文本处理基础

1) 正则表达式

自然语言处理过程中面临大量的文本处理工作，如词干提取、网页正文抽取、分词、断句、文本过滤、模式匹配等任务，而正则表达式往往是首选的文本预处理工具。

现在主流的编程语言对正则表达式都有较好的支持，如Grep、Awk、Sed、Python、Perl、Java、C/C++ (推荐re2 (<http://code.google.com/p/re2/>))等。

注：课程中给出的正则表达式语法和示例在此略去

2) 分词

• 词典规模

同一词条可能存在不同的时态、变形，那么给定文本语料库，如何确定词典规模呢？



首先定义两个变量Type和Token：

-Type：词典中的元素，即独立词条

-Token：词典中独立词条在文本中的每次出现

如果定义 $N = \text{number of tokens}$ 和 $V = \text{vocabulary} = \text{set of types}$ ， $|V|$ is the size of the vocabulary，那么根据Church and Gale (1990) (http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CGUQFjAA&url=http%3A%2F%2Facl.ldc.upenn.edu%2FH%2FH90%2FH90-1056.pdf&ei=guWIT4rCFOWZ2QW3mbymAg&usg=AFQjCNGHW9r2i_vG1r2x5OeJEKWYyBwhAw&sig2=AVNUU5dffejqy1oZyDterg)的研究工作可知： $|V| > O(N^{1/2})$ ，验证如下：

http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CGUQFjAA&url=http%3A%2F%2Facl.ldc.upenn.edu%2FH%2FH90%2FH90-1056.pdf&ei=guWIT4rCFOWZ2QW3mbymAg&usg=AFQjCNGHW9r2i_vG1r2x5OeJEKWYyBwhAw&sig2=AVNUU5dffejqy1oZyDterg的研究工作可知： $|V| > O(N^{1/2})$ ，验证如下：

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>

先验和后验概率 (<http://blog.csdn.net/jingtingxu369/article/details/78006999>)



移民澳大利亚

相关推荐

斯坦福大学自然语言处理基础 (Basic Text Processing) ” (<http://blog.csdn.net/kyiy/article/details/37565321>)



二手豪华车

Stanford第二课“文本处理基础 (Basic Text Processing) ” (<http://blog.csdn.net/kyiy/article/details/37565321>)

Stanford自然语言处理笔记1 -basic text processing (<http://blog.csdn.net/u011954647/article/details/50663018>)

Spark2.0特征提取与转换选择之182特征选择文章处理/52432自然语言处理



日语常用语

环氧自流平地面

自流平地面

零基础学习手绘



内容举报



返回顶部



登录

注册



	Tokens = N	Types = V
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million



0

分词算法

任务：统计给定文本文件（如shake.txt）中词频分布，子任务：分词，频率统计，实现如下：



```
tr 'A-Z' 'a-z' < shakes.txt
```

Merging upper and lower case



```
| tr -sc 'A-Za-z' '\n'
```

Change all non-alpha to newlines

```
| sort
```

Sort in alphabetical order

```
| uniq -c
```

Count the frequency

```
| sort -n -r
```

Sorting the counts

以上实现将非字母字符作为token分隔符作为简单的分词器实现，但是，这难免存在许多不合理之处，如下面的例子：

- Finland's capital -> Finland Finlands Finland's ?
- what're, I'm, isn't -> What are, I am, is not
- Hewlett-Packard -> Hewlett Packard ?
- state-of-the-art -> state of the art ?
- Lowercase -> lower-case lowercase lower case ?

San Francisco -> one token or two?

- m.p.h., PhD. -> ??

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/>)





¥268.00/套

移民澳大利亚

二手豪华车

他的热门文章

opencv的安装 (<http://blog.csdn.net/jingtingxu369/article/details/52986177>)

1212

①协方差、相关系数（皮尔逊相关系数），等同于：内积、余弦值。(<http://blog.csdn.net/jingtingxu369/article/details/53843159>)

1188

caffe权值可视化,特征可视化,网络模型可视化 (<http://blog.csdn.net/jingtingxu369/article/details/52997586>)

1047

机器学习中overfitting的理解 (<http://blog.csdn.net/jingtingxu369/article/details/49805385>)

940

caffe中运行有关.py的程序需要安装pycaffe, 安装如下 (<http://blog.csdn.net/jingtingxu369/article/details/49805385>)



内容举报



返回顶部



上面的方法对英语这种包含固定分隔符的语言行之有效，对于汉语、日语、德语等文本则不再适用，所以需要专门的分词技术。其中，最简单、最常用，甚至是最有效的方法就是最大匹配法（Maximum Matching），它是一种基于贪心思想的切词策略，主要包括正向最大匹配法（Forward Maximum Matching, FMM）、逆向最大匹配法（Backward Maximum Matching）与双向最大匹配法（Bi-direction Maximum Matching, BMM）。

以FMM中文分词为例，步骤如下：

- 选取包含N(N通常取6~8)个汉字的字符串作为最大字符串；
- 把最大字符串与词典中的单词条目相匹配（词典通常使用Double Array Trie (http://code.google.com/p/darts-clone/)组织）；
- 如果不能匹配，就去掉最后一个汉字继续匹配，直到在词典中找到相应的词条为止。

例如：句子“莎拉波娃现在居住在美国东南部的佛罗里达”的分词结果是“莎拉波娃/ 现在/ 居住/ 在/ 美国/ 东南部/ 的/ 佛罗里达”。

另外，使用较多的分词方法有最少分词法、最短路径法、最大概率法（或词网格法，大规模语料库+HMM/HHMM）、字标注法等。

分词难点

-切分歧义：主要包括交集型歧义和覆盖型歧义，在汉语书面文本中占比并不大，而且一般都可以通过规则或建立词表解决。

-未登录词：就是未在词典中记录的人名（中、外）、地名、机构名、新词、缩略语等，构成了中文自然语言处理永恒的难点。常见的解决方法有互信息、语言模型，以及基于最大熵或隐马尔科夫模型的统计分类方法。

3) 文本归一化

主要包括大小写转换、词干提取、繁简转换等问题。

4) 断句

句子一般分为大句和小句，大句一般由“！”，“。”，“；”，“\”，“？”等分割，可以表达完整的含义，小句一般由“，”分割，起停顿作用，需要上下文搭配表达特定的语义。

中文断句通常使用正则表达式将文本按照有分割意义的标点符号(如句号)分开即可，而对于英文文本，由于英文句号“.”在多种场景下被使用，如缩写“Inc.”、“Dr.”、“0.02%”、“4.3”等，无法通过简单的正则表达式处理，为了识别英文句子边界，课程中给出了一种基于决策树（Decision Tree）的分类方法，如下图所示：

xu369/article/details/53084336 (https://passport.csdn.net/a

914



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册



(https://passport.csdn.net/a

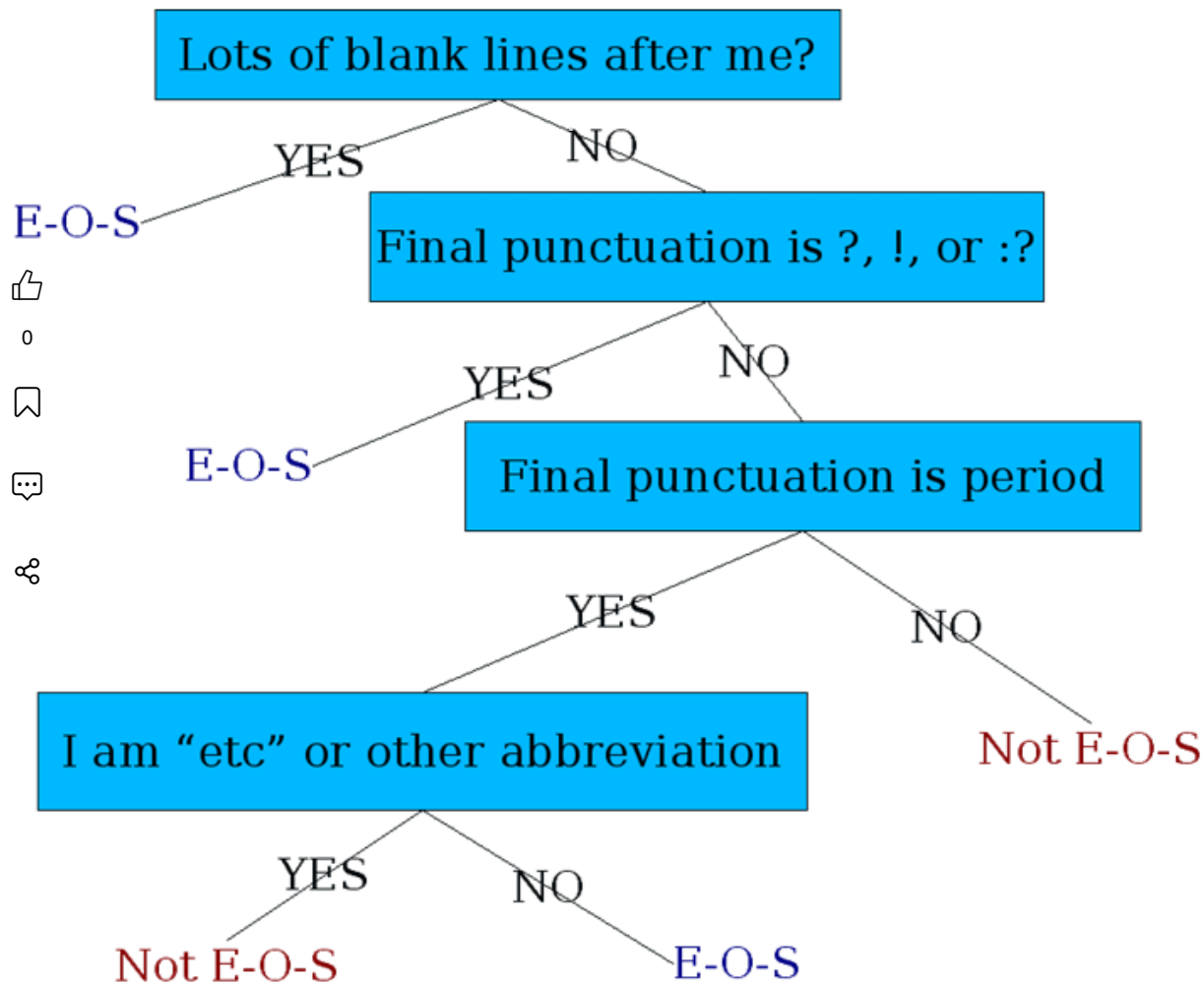


移民澳大利亚



二手豪华车

广告



此方法的核心就是如何选取有效的特征？如句号前后的单词是否大写开头、是否为缩略词、前后是否存在数字、句号前的单词长度、句号前后的单词在语料库中作为句子边界的概率等等。当然，你也可以基于上述特征采用其他分类器解决断句问题，如逻辑回归（Logistic regression）、支持向量机（Support Vector Machine）、神经网络（Neural Nets）等。

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册



⚠
内容举报

⬆
TOP
返回顶部





0



相关文章推荐



斯坦福大学自然语言处理第二课“文本处理基础 (Basic Text Processing) ” (<http://blog.csdn...>)

文本处理基础1.正则表达式(Regular Expressions)正则表达式是重要的文本预处理工具。以下截取了部分正则写法：2.分词 (Word tokenization) 我们在...

 IOThouzhuo (<http://blog.csdn.net/IOThouzhuo>) 2015年08月26日 18:47  1477

stanford第二课“文本处理基础 (Basic Text Processing) ” (<http://blog.csdn.net/fkyyly/artic...>)

一、课程介绍 斯坦福大学于2012年3月在Coursera启动了在线自然语言处理课程，由NLP领域大牛Dan Jurafsky 和 Chirs Manning教授授课： <https://cla...>

 fkyyly (<http://blog.csdn.net/fkyyly>) 2014年07月08日 15:11  848



广告

惊呆了！微博和阿里背后的数据库有多厉害？

想不到！数据库作为最关键的基础设施，渗透技术领域的方方面面，我阿里和微博的师哥们是这么分享的...

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

[登录](#)[注册](#)<https://passport.csdn.net/a>

移民澳大利亚



二手豪华车

广告



内容举报



返回顶部

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjTzrjb0IZ0qnfK9uYzP1nsrjD10Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Yznyc3nWn3uyRvrHRdryN90AwY5HDdnHn3rjb3nWc0IgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEpZNGXy-jULNzTvRETVnzpyN1gww-IA7GUatLPjqdIAdxTvqdThP-5yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpYfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPjfLnj0)

Stanford自然语言处理笔记1 -basic text processing (http://blog.csdn.net/u011954647/artic...

//过年归来，虽然还想玩，但是还是要学习了，上午机器学习技法，下午nlp吧。= //最近在看公主的男人，我家厚厚好帅=。
=，实在是太呆萌了。 开始 第一部分直接讲的正则匹配=。= We和we...

u011954647 (http://blog.csdn.net/u011954647) 2016年02月14日 17:29 493

Spark2.0 特征提取、转换、选择之二：特征选择、文本处理，以中文自然语言处理(情感分类为...

Spark2.0文本特征提取

qq_34531825 (http://blog.csdn.net/qq_34531825) 2016年09月04日 11:15 2778



Foundations_of_Statistical_Natural_Language_Processing.pdf统计自然...

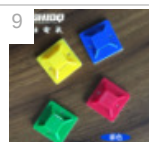
(http://download.csdn.net/detail/qq_34531825/9444444) 2010年06月22日 11:19 7.52MB 下载



180.00/件
订做精图 网络箱 ONU
箱 综合箱 网络配线箱



30.00/件
12芯 24芯 ST SC FC
机架式光纤盒，光缆终



18.00/包
【批发】塑料吸盘定位
片25*25 不干胶自粘式



TPL (Text Processing Library文本处理库) 下载 (http://download.csdn.n...

(http://download.csdn.net/detail/qq_34531825/9444444) 2009年06月25日 15:02 457KB 下载

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册




内容举报



返回顶部

Rich Text Processing富文本处理 (<http://blog.csdn.net/xuguangsoft/article/details/8579932>)

Scribe框架提供一系列读和控制富文本文档的类。Qt4提供像QTextDocument类，他能够为开发提供创建和修改结构的富文本文档。文档内的信息通过两个补充的接口存取：1. 基于光标的...

 xuguangsoft (<http://blog.csdn.net/xuguangsoft>) 2013年02月13日 14:11 8084



0



Text Processing In Python (Python文本处理) (<http://download.csdn.net/d...>)

(<http://download.csdn.net/d...>) 2009年12月21日 07:35 1.35MB [下载](#)

统计自然语言处理基础学习笔记(8)——文本分析 (<http://blog.csdn.net/dqjyong/article/details...>)

自然语言处理的目的是为了更好的分析人类语言，让机器能够理解人类的语言。随着互联网的兴起，人们越来越多的参与网络社区活动，人们在网络社区发言的机会越来越多，文本分析的需求也越来越迫切。而依靠人工去分析这...

 dqjyong (<http://blog.csdn.net/dqjyong>) 2014年03月02日 22:05 3647

用Python进行自然语言处理-1. Language Processing and Python (<http://blog.csdn.net/reb...>)

《用Python进行自然语言处理》是一本结合了自然语言处理和Python知识的入门书籍，现在书籍正在出第二版，预计2016年完成。第二版是与Python 3配套的，很多地方都要修改。附上书籍原地址链接...

 rebellion51 (<http://blog.csdn.net/rebellion51>) 2015年07月27日 15:41 581



自然语言处理综论英文版Speech and Language Processing (<http://downlo...>)

(<http://download.csdn.net/d...>) 2012年04月20日 18:29 36.23MB [下载](#)

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！
Natural Language Processing with Python PYTHON自然语言处理中文翻...



(<https://passport.csdn.net/a>)



移民澳大利亚



二手豪华车

广告



内容举报




返回顶部

登录



注册



 [/http://download.csdn.net/detail/garfielder007/10111111](#) 2014年08月25日 16:09 3.94MB [下载](#)

自然语言处理List of 25+ Natural Language Processing APIs ([http://blog.csdn.net/Garfield...](http://blog.csdn.net/GarfieldEr007)


Natural Language Processing API Note: Check out our latest API collections page for the lis...

 GarfieldEr007 (<http://blog.csdn.net/GarfieldEr007>) 2016年04月20日 13:50  988

0




网易新闻语料库 文本分类 自然语言处理 ([http://download.csdn.net/downlo...](http://download.csdn.net/download/garfielder007/10111111)

 [/http://download.csdn.net/detail/garfielder007/10111111](#) 2013年12月11日 00:49 37.74MB [下载](#)



自然语言处理语料库新闻文本凤凰新闻第一部分 ([http://download.csdn.net/...](http://download.csdn.net/detail/garfielder007/10111111)

 [/http://download.csdn.net/detail/garfielder007/10111111](#) 2010年05月08日 13:57 19.28MB [下载](#)

MIT自然语言处理第二讲：单词计数（第一、二部分）([http://blog.csdn.net/xiaopihaierletian...](http://blog.csdn.net/xiaopihaierletian)

MIT自然语言处理第二讲：单词计数（第一部分）自然语言处理：单词计数 Natural Language Processing: (Simple) Word Co unt...

 xiaopihaierletian (<http://blog.csdn.net/xiaopihaierletian>) 2017年04月07日 10:03  76



Python 自然语言处理 第二版 20170304 ([http://download.csdn.net/downlo...](http://download.csdn.net/download/garfielder007/10111111)

[/http://download.csdn.net/detail/garfielder007/10111111](#) 2017年03月04日 15:02 14.61MB [下载](#)

宗成庆 统计自然语言处理（第二版）.pdf ([http://download.csdn.net/downl...](http://download.csdn.net/download/garfielder007/10111111)

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/a>



移民澳大利亚



二手豪华车

广告



内容举报



返回顶部

登录

注册



2017年11月15日 09:22 20.47MB

下载

MIT自然语言处理第一讲：简介和概述（第二部分）(http://blog.csdn.net/xiaopihaierletian/a...

自然语言处理：背景和概述 Natural Language Processing:Background and Overview 作者：Regina Barzilay (MIT,EECS De par...



xiaopihaierletian (http://blog.csdn.net/xiaopihaierletian)

2017年04月07日 09:59

109



0



统计自然语言处理 宗成庆（第二版）(http://download.csdn.net/download/...

(http://download.csdn.net/download/...

2017年12月16日 23:08

55.65MB

下载



(https://passport.csdn.net/a



移民澳大利亚



二手豪华车

广告



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

[登录](#)[注册](#)