



博客 (<http://blog.csdn.net?ref=toolbar>)

学院 (<http://edu.csdn.net?ref=toolbar>)

下载 (<http://download.csdn.net?ref=toolbar>)

GitChat (<http://gitbook.cn/?ref=csdn>)

1

写博客

发Chat

登录 (<https://passport.csdn.net/account/login?ref=toolbar>) 注册 (<https://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister>)

字符串相似性的几种度量方法

原创 2016年11月09日 21:58:57

标签：字符串相似性 (<http://so.csdn.net/so/search/s.do?q=字符串相似性&t=blog>)

3318

无论是做科学研究，还是工程项目，我们总是会碰上要比较字符串的相似性，比如拼写纠错、文本去重、上下文相似性等。度量的方法有很多，到底使用哪一种方法来计算相似性，这就需要我们根据情况选择合适的方法来计算。这里把几种常用到的度量字符串相似性的方法罗列一下，仅供参考，欢迎大家补充指正。

1、余弦相似性 (cosine similarity) (https://en.wikipedia.org/wiki/Cosine_similarity)
余弦相似性大家都非常熟悉，它是定义在向量空间模型 (VSM) 中的。它的定义如下：

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

其中，A,B为向量中间中的两个向量。
在使用它来做字符串相似性度量的时候，需要先将字符串向量化，通常使用词袋模型 (BOW) 来向量化。举个例子如下：

String1 = "apple"
String2 = "app"

则词包为 { ' a ' , ' e ' , ' l ' , ' p ' } ，若使用0,1判断元素是否在词包中，字符串1、2可以转化为：

StringA = [1111]
StringB = [1001]

那么，根据余弦公式，可以计算字符串相似性为：0.707。

2、欧氏距离 (Euclidean distance) (https://en.wikipedia.org/wiki/Euclidean_distance)
欧氏距离大家非常熟悉，定义在向量空间模型中，计算使用欧氏距离公式：

$$\begin{aligned} d(\mathbf{p}, \mathbf{q}) &= d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \end{aligned}$$



火食三刀 (<http://blog.cs...>)

+ 关注

(http://blog.csdn.net/shijing_0214)

码云
0
(<https://g...>)
utm_sour

原创	粉丝	喜欢
50	100	4



人脸识别技术



他的最新文章

更多文章 (http://blog.csdn.net/shijing_0214)

理解ResNet (http://blog.csdn.net/shijing_0214/article/details/78475372)

windows下使用tensorflow简单实现CNN (http://blog.csdn.net/shijing_0214/article/details/76167012)

为什么特征独立型的模型遇到高度相关特征效果会不好？ (http://blog.csdn.net/shijing_0214/article/details/75864342)

序列模型中的注意力机制 (http://blog.csdn.net/shijing_0214/article/details/75194103)

六大排序算法 (插冒归堆选择快排) (http://blog.csdn.net/shijing_0214/article/details/73033332)

文章分类

深入理解Java虚拟机 (http://b...)	3篇
机器学习 (http://blog.csdn.n...)	20篇
自然语言处理 (http://blog.cs...)	8篇
神经网络 (http://blog.csdn.n...)	5篇
数据挖掘与推荐 (http://blog....)	4篇

展开

3、编辑距离（edit distance）(https://en.wikipedia.org/wiki/Edit_distance)

编辑距离，有的地方也会称为Levenshtein距离，表示从一个字符串转化为另一个字符串所需要的最少编辑次数，这里的编辑是指将字符串中的一个字符替换成另一个字符，或者插入删除字符。例如上例String1通过删除‘l’与‘e’转化为String2，所以其最小编辑次数为2。
编辑距离的核心就是如何计算出一对字符串间的最小编辑次数，考虑到问题的特点，我们可以使用动态规

划的思想来计算其最小编辑次数，根据维基百科：两个字符串的编辑距离递归计算公式如下：

$$\begin{aligned} d_{i0} &= \sum_{k=1}^i w_{\text{del}}(b_k), & \text{for } 1 \leq i \leq m \\ d_{0j} &= \sum_{k=1}^j w_{\text{ins}}(a_k), & \text{for } 1 \leq j \leq n \\ d_{ij} &= \begin{cases} d_{i-1,j-1} & \text{for } a_j = b_i \\ \min \begin{cases} d_{i-1,j} + w_{\text{del}}(b_i) \\ d_{i,j-1} + w_{\text{ins}}(a_j) \\ d_{i-1,j-1} + w_{\text{sub}}(a_j, b_i) \end{cases} & \text{for } a_j \neq b_i \end{cases} & \text{for } 1 \leq i \leq m, 1 \leq j \leq n. \end{aligned}$$

其中，w表示增删改三种操作的权重，一般定义为：

$$w = \begin{cases} 1, & \text{若有操作} \\ 0, & \text{无操作} \end{cases}$$

$d_{i0} = i$ 表示从 $b' = b_1 \cdots b_i$ 删除为空的编辑次数； $d_{0j} = j$ 表示从空插入成 $a' = a_1 \cdots a_j$ 所需的编辑次数； d_{ij} 则是对动态规划中分解子问题的过程。
仍以（1）中的两个字符串为例：

则编辑距离 $d_{53} = \min \begin{cases} d_{43} + 5, & \text{删除操作} \\ d_{52} + 3, & \text{插入操作，继续通过不断递归可以得出其编辑距离。} \\ d_{42} + 0, & \text{替换操作} \end{cases}$

4、海明距离（hamming distance）(https://en.wikipedia.org/wiki/Hamming_distance)

海明距离用于表示两个等长字符串对应位置不同字符的总个数，也即把一个字符串换成另一个字符串所需要的替换操作次数。根据定义，可以把海明距离理解为编辑距离的一种特殊情况，即只计算等长情况下替换操作的编辑次数。举个例子来讲，字符串“bob”与“pom”的海明距离为2，因为需要至少两次的替换操作两个字符串才能一致。海明距离较常用与二进制串上的操作，如对编码进行检错与纠错。在计算长字符串的相似性时可以通过hash函数将字符串映射成定长二进制串再利用海明距离来计算相似性。
海明距离的计算比较简单，通过一个循环来比较对应位置的字符是否相同即可。

5、Dice 距离 (https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient)

Dice距离用于度量两个集合的相似性，因为可以把字符串理解为一种集合，因此Dice距离也会用于度量字符串的相似性。此外，Dice系数的一个非常著名的使用即实验性能评测的F1值。Dice系数定义如下：

$$QS = \frac{2|X \cap Y|}{|X| + |Y|}$$

其中，X,Y表示两个集合，分子表示两个集合的相交操作后的长度，分母表示两个集合长度之和。以（1）中的例子来讲的话， $dice_{12} = \frac{2 \times 3}{5+3} = 0.75$ 。若集合表示成向量的话，计算可以定义为：

$$s_v = \frac{2|A \cdot B|}{|A|^2 + |B|^2}$$

文章存档

2017年11月 (http://blog.csdn.net/shijing_0214/article/details/51757564)	1篇
2017年7月 (http://blog.csdn.net/shijing_0214/article/details/51757564)	3篇
2017年6月 (http://blog.csdn.net/shijing_0214/article/details/51757564)	1篇
2017年5月 (http://blog.csdn.net/shijing_0214/article/details/51757564)	3篇
2017年4月 (http://blog.csdn.net/shijing_0214/article/details/51757564)	1篇

展开

他的热门文章

- 几种范数的简单介绍 (http://blog.csdn.net/shijing_0214/article/details/51757564) 25160
- Unicode与UTF-8的差别 (http://blog.csdn.net/shijing_0214/article/details/51757564) 19281
- 简单理解LSTM神经网络 (http://blog.csdn.net/shijing_0214/article/details/52081301) 12691
- 理解支持向量机（二）核函数 (http://blog.csdn.net/shijing_0214/article/details/51000845) 11760
- 理解数学空间，从距离到希尔伯特空间 (http://blog.csdn.net/shijing_0214/article/details/51052208) 10349

怎样去除黑眼圈



联系我们

- 网站客服 微博客服 (http://wpa.qq.com/msgrd?v=3&uin=2431299880&site=qq&chat=1)
- (http://e.weibo.com/csdnsupport/)
- webmaster@csdn.net (mailto:webmaster@csdn.net)
- 400-660-0108

京ICP证09002463号
(http://www.miibeian.gov.cn/)

其中，A,B表示两个向量。

6、Jaccard distance (https://en.wikipedia.org/wiki/Jaccard_index)

杰卡德系数的定义如下，

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

可以看出与Dice系数的定义比较相似，分子部分是个两倍关系，Dice系数的分母比Jaccard系数的分母多减去了一项分子，即 $|A \cap B|$ 。

Jaccard与Dice之间具有一种转化关系：

$$J = \frac{D}{2-D},$$

或：

$$D = \frac{2J}{1+J}$$

7、J-W距离 (Jaro-Winkler distance)

(https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance)

J-W距离也常用来度量两个字符串的相似性，它实际上 Jaro distance的一种变种。Jaro distance距离属于编辑距离的一类，被用于记录链接领域来将异构数据源中的records链接到同义实体中，也可以用于拼写纠错。Jaro distance定义如下：

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases}$$

其中，m是两个字符串匹配上的字符数目，t是字符中换位数目的一半，即若在字符串的第i位出现了a,b，在第j位又出现了b,a，则表示两者出现了换位。举个例子来讲：

$s_1 = MARTHA$

$s_2 = MARHTA$

则 $m = 6, |s_1| = 6, |s_2| = 6, T/H$ 和 H/T 属于两对换位字符对，故 $t = \frac{1+1}{2} = 1$

代入公式可得： $J_{1,2} = 0.944$ 。一般定义当J值不大于 $\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1$ 时两个字符串被匹配上。

有了Jaro distance，我们定义J-W距离：

$$d_w = \begin{cases} d_j & \text{if } d_j < b_t \\ d_j + (\ell p(1 - d_j)) & \text{otherwise} \end{cases}$$

其中， d_j 即为Jaro距离； ℓ 是字符串的起始最大公共前缀，最大不超过4； p 为一个缩放因子，用于对J进行调整，避免 d_w 超出1，一般设为0.1； b_t 为boost threshold，当值超过该值时激发Jaro距离为J-W距离，该值一般设为0.7。

仍以上面的两个字符串为例， $d_j = 0.944 > 0.7, \ell = 3, p = 0.1$ ，代入公式可算出 $d_w = 0.961$ 。

关于

(<http://www.csdn.net/company/about.h>

招聘

(<http://www.csdn.net/company/recruit.h>

广告服务

(<http://www.csdn.net/company/marketin>

阿里云

Copyright © 1999-2018

CSDN.NET, All Rights Reserved



计算字符串的相似度-两种解法 zzran 2012年12月09日 13:34 32316

1 一直不理解，为什么要计算两个字符串的相似度呢。什么叫做两个字符串的相似度。经常看别人的博客，碰到比较牛的人，然后就翻了翻，终于找到了比较全面的答案和为什么要计算字符串相似度的解释。因为搜索引擎要把通过...
(http://blog.csdn.net/zzran/article/details/8274735)



字符串相似度算法介绍(整理) dongle2001 2007年01月02日 12:14 23347

最近在做这方面的应用，把我找到的资料贴出来，有需要的人可以参考参考。1．编辑距离（Levenshtein Distance）编辑距离就是用来计算从原串（s）转换到目标串(t)所需要的最少的插入，删除和...
(http://blog.csdn.net/dongle2001/article/details/1472235)

【揭秘】程序员升职加薪的捷径来了！

在岗5年，总想着闲下来的时候应该如何安排自己的程序人生呢？无意中看到这个！眼睛亮了..



(http://www.baidu.com/cb.php?c=IgF_pyfqHmknjT3P160IZ0qnfK9ujYzP1nsrjDz0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y4PvFhPHKbrHIWnAfzrHuB0AwY5HDdnHRdP1nvrjD0IgF_5y9YIZ0IQzqBTLn8mLPbUB48ugfEUiqYULKGmzq-uZNxug99UHqdIAdxTvqdThP-5yF_UvTkn0KzujY1n0KBuHYs0ZKz5H00Iy-b5HDdP1f1PWD0Uv-b5HDzrH63nHf0mv-b5HTzPWb1n6KEIv3qn0KsXHYznjm0mLFW5H63P0)

字符串相似度算法(编辑距离Levenshtein Distance)

什么是Levenshtein 编辑距离（Edit Distance），最先是由俄国科学家Vladimir Levenshtein在1965年发明，用他的名字命名，又称Levenshtein距离。是...

chndata 2015年01月09日 11:38 2757

(http://blog.csdn.net/chndata/article/details/42552971)

字符串相似度算法及应用 mingspy 2012年05月30日 13:14 4343

Levenshtein Distance 简介 字符串相似度的算法还是比较经典的DP算法，看到有两篇文章写的比较不错，他们的介绍也非常详细，值得学习。文章地址如下: 文章1 http://blog...
(http://blog.csdn.net/mingspy/article/details/7615855)

算法系列之四：字符串的相似度 orbit 2011年07月31日 23:39 22367

算法系列之四：字符串的相似度 我们把两个字符串的相似度定义为：将一个字符串转换成另外一个字符串的代价（转换的方法可能不唯一），转换的代价越高则说明两个字符串的相似度越低。比如两...
(http://blog.csdn.net/orbit/article/details/6649322)

【机器学习】恭喜，免费试听机器学习课程





立即试听

图像分割结果的评估 zhuason 2016年11月01日 00:15 6044

我们在用一个算法对一幅图像进行分割之后，总会面临这样一个问题，分割的结果到底好不好。用眼睛可以看出好坏，但这只是主观的好坏，如何量化的对分割的结果进行评价呢，这是这篇文章我要讨论的主题。 我查阅过...
(http://blog.csdn.net/zhuason/article/details/52989091)

Dice's coefficient

 gjk0223 2008年04月22日 14:21  4416



Dices coefficient (also known as the Dice coefficient) is a similarity measure related to the Jaccard...

(<http://blog.csdn.net/gjk0223/article/details/2314844>)

算法之美——求解 字符串间最短距离（动态规划）

1 算法之美——求解 字符串间最短距离（动态规划） 分类： 算法 动态规划 2012-09-04 18:20 1796人阅读 评论(2) 收藏 举报 distance string 算法 in...





 jfkidear 2014年09月07日 11:44  2911



(<http://blog.csdn.net/jfkidear/article/details/39118847>)



字符串的距离

 ACdreamers 2013年11月12日 14:34  3864

题目：<http://wikioi.com/problem/2180/> 题意：设有字符串X，我们称在X的头尾及中间插入任意多个空格后构成的新字符串为X的扩展串，如字符串X为“abcbcd”，则 ...

(<http://blog.csdn.net/ACdreamers/article/details/15502107>)



计算两个字符串的距离

 u014482079 2014年09月10日 16:36  712

题目描述：计算 对于不同的字符串，希望能够有办法判断其相似程度。定义了如下方法来把两个不同的字符串变得相同，具体的操作方法为：1. 修改一个字符（如把“a”替换为“b”）2. 增...

(<http://blog.csdn.net/u014482079/article/details/39181947>)



华为OJ（计算字符串的距离）

 yiqiawangxi 2015年08月22日 19:58  1569

描述 Levenshtein 距离，又称编辑距离，指的是两个字符串之间，由一个转换成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。编辑距...

(<http://blog.csdn.net/yiqiawangxi/article/details/47862857>)



字符串间最短距离（动态规划）

 jie1991liu 2013年04月09日 17:46  4468

Minimum Edit Distance 问题 解法一：对于不同的字符串，判断其相似度。 定义了一套操作方法来把两个不相同的字符串变得相同，具体的操作方法为： ...

(<http://blog.csdn.net/jie1991liu/article/details/8778893>)

《编程之美》——计算字符串的相似度



 zengzhen_CSDN 2015年11月03日 17:07  1415

问题：许多程序会大量使用字符串。对于不同的字符串，我们希望能够有办法判断其相似程度。我们定义一套操作方法来把两个不相同的字符串变得相同，具体的操作方法为：1.修改一个字符（如把“a”替换为“b”）...

(http://blog.csdn.net/zengzhen_CSDN/article/details/49618895)



18种和“距离(distance)”、“相似度(similarity)”相关的量的小结

在计算机人工智能领域，距离(distance)、相似度(similarity)是经常出现的基本概念，它们在自然语言处理、计算机视觉等子领域有重要的应用，而这些概念又大多源于数学领域的度量(metric...)...

 solomonlangrui 2015年08月12日 23:16  7994

(<http://blog.csdn.net/solomonlangrui/article/details/47454805>)



Jaro-Winkler Distance 浅析

 chaoswork 2010年04月15日 17:56  13637

这是一种计算两个字符串之间相似度的方法，想必都听过Edit Distance, Jaro-Winkler Distance 是Jaro Distance的一个扩展，而Jaro Distance...

(<http://blog.csdn.net/chaoswork/article/details/5489877>)



字符串相似算法-Jaro-Winkler Distance


 Gatherfly 2015年12月08日 12:41  514

Jaro-Winkler Distance 算法 这是一种计算两个字符串之间相似度的方法，想必都听过Edit Distance,Jaro-inkler Distance 是Jaro Distance...

(<http://blog.csdn.net/Gatherfly/article/details/50217197>)



字符串相似度算法


 chinesesword 2012年06月07日 10:01  4093

 原文:<http://blog.csdn.net/guffey/article/details/6750494> 2011-09-05 17:30 74人阅读 评论(0) 收藏 举报 ...

 1 (<http://blog.csdn.net/chinesesword/article/details/7640787>)



python文本相似度之距离计算详细介绍

 qq_27713281 2017年05月24日 14:45  773

 编辑距离 编辑距离（ Edit Distance ），又称Levenshtein距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除...

(http://blog.csdn.net/qq_27713281/article/details/72676282)



华为OJ（计算字符串的相似度）

 yiqiawangxi 2015年08月15日 17:43  1273

题目：计算字符串的相似度 描述 对于不同的字符串，我们希望能有办法判断相似程度，我们定义了一套操作方法来把两个不相同的字符串变得相同，具体的操作方法如下：1 修改一个字符，如把“a”替换为...

(<http://blog.csdn.net/yiqiawangxi/article/details/47683871>)

计算字符串的相似度-两种解法

 jfkidear 2016年10月25日 22:57  352

计算字符串的相似度-两种解法 2012-12-09 13:34 15769人阅读 评论(0) 收藏 举报 版权声明：本文为博主原创文章，未经博主允许不得转载。 一...

(<http://blog.csdn.net/jfkidear/article/details/52928471>)