

<u>28</u>

...

快速入门

首先,让我们导入gensim并创建一个小小的语料库,其中有9篇文档和12个属性[1] (http://www.cs.bham.ac.uk/~pxt/IDA/lsa_ind.pdf):

```
[python]
      >>> from gensim import corpora, models, similarities
      >>>
      >>> corpus = [[(0, 1.0), (1, 1.0), (2, 1.0)],
      >>>
                    [(2, 1.0), (3, 1.0), (4, 1.0), (5, 1.0), (6, 1.0), (8, 1.0)],
      >>>
                    [(1, 1.0), (3, 1.0), (4, 1.0), (7, 1.0)],
                    [(0, 1.0), (4, 2.0), (7, 1.0)],
      >>>
                    [(3, 1.0), (5, 1.0), (6, 1.0)],
                   Γ(9, 1.0)],
      >>>
                   [(9, 1.0), (10, 1.0)],
10.
      >>>
                    [(9, 1.0), (10, 1.0), (11, 1.0)],
11. >>>
                    [(8, 1.0), (10, 1.0), (11, 1.0)]]
```

语料库是一个简单的对象,我们能够通过不断迭代从中取出用稀疏向量代表的文档。如果你不熟悉向量空间模型,我们将会在下一个教程(语料库与向量空间)(http://blog.csdn.net/questionfish/article/details/46739207)填平原始字符串、语料库、稀疏向量之间的鸡沟。如果你熟悉向量空间模型,你可能也知道如何将你的文档解析、转锋为向量,并且也知道这些转换对接下来的应用的会产生什么影响。

[plain]

1. 作者注:在这个例子中我们的整个语料作为一个Python List被存在内存中。但是,语料库接口仅仅规定一个语料库必须支持迭 代取出其文档。对于特别大的语料库,最好榜整个语料库存在硬盘上,并且按序一次取出一篇文档。所有的操作和特换通过一种 语料库人小无关的方式深度,内存依赖效应。

接下来,让我们初始化一个转换

```
[python]
1. >>> tfidf = models.TfidfModel(corpus)
```

转换就是将文档的一种向量表示方法转换为另一种 (, 以便我们从特定的角度更好地分析数据)

```
[python]

1. >>> vec = [(0, 1), (4, 1)]

2. >>> print(rtioff/vec)]

3. [(0, 0.807524), (4, 0.5898342)]
```

在此,我们使用了Tf-ldf (http://www.ruamyifeng.com/blog/2013/03/tf-idf.html),这是一种简单的转换。它要求输入的文档用带有词频的词袋的方法 表示,可以用来降低常用词的权量(相对地提高了罕见词的权量)。它还会把结果向量的长度调整为单位长度(指欧几里得范数)。 转化方法详情,请看教程《主题与转换》(http://blog.csdn.net/questionfish/article/details/46742403)

为了将整个语料库通过Tf-idf转化并索引,以便相似度查询,需要做如下准备

```
[python]
1. | >>> index = similarities.SparseMatrixSimilarity(tfidf[corpus], num_features=12)
```

为了查询我们需要的向量vec相对于其他所有文档的相似度,需要

```
[python]

1. >>> sims = index[tfidf[vec]]

2. >>> print(list(enumerate(sims)))

3. [(0, 0.4662244), (1, 0.19139354), (2, 0.24600551), (3, 0.82094586), (4, 0.0), (5, 0.0), (6, 0.0)
```

如何解释这些输出呢?文档0(第1个文档)与vec的相似度为0.466=46.6%,第二个文档与vec的相似度为



达人课



Adx TyqqTh

Adx TyqqTh

Adx TyqqTh

THE SHAWOW RASKS BY SHEW SHAWOW RESERVE TO THE SHAWOW RESERVE TO THE WORLD TO THE SHAWOW RESERVE TO THE SHAWOW

Blackfig1RLPWT0TAg15H 自由ANNS在1944和ANN(中)——快速入 OKANACYANACAGAKANACAGAISh/articl

DS/BANGEON/ZEON/OGABLOO LIBERTO BOOTIZHAPE CANGREP

5HDznHNImhkEusKzujY 6ANSが単99野発配等中型) — 相似度 電地でいいでは、 1年の19年7年の1

WThnqPH04PW6) 回 5588 NumPy、SciPy 等Python包在Windows下 的whl安装包下载 (http://blog.csdn.net/qu

estionfish/article/details/46963867)

Gensim官方教程翻译(三)——主题与 转换(Topics and Transformations) (htt p://blog.csdn.net/questionfish/article/detai ls/46742403)

□ 5154

CC 5345



企 返回顶部

介容举报 介容学报 企 返回顶部

ß

19.1,依次类推。 因此,根据Tfidf文档表示方法和余弦相似度方法,与我们的查询文档vec最相似的文档为 3号文档,相似度达到82.1%。注意:4-8号文档与vec没有任何公共的属性,因此相似度为0.0。
了解详细请看《相似度查询》 (http://radimrehurek.com/gensim/tut3.html)教程。
[1] (http://radimrehurek.com/gensim/tutorial.html#id1) This is the same corpus as used in Deerwester et al. (1990): Indexing by Latent Semantic Analysis (http://www.cs.bham.ac.uk/-pxt/IDA/Isa_ind.pdf), Table 2.
发表你的评论 (http://my.csdn.net/weixin 35068028)
相关文章推荐
Gensim Word2vec 使用教程 (http://blog.csdn.net/Star_Bob/article/details/47808499)
本文主要基于Radim Rehurek的Word2vec Tutorial.**准备输入**Gensim的word2vec的输入是句子的序列. 每个句子是一个单 词列表代码块例如:>>># impor
Star_Bob (http://blog.csdn.net/Star_Bob) 2015年08月20日 15:26 □25271
python gensim训练word2vec及相关函数与功能理解 (http://blog.csdn.net/sinat_26917383
一、gensim介绍gensim是一款强大的自然语言处理工具,里面包括N多常见模型: - 基本的语料处理工具 - LSI - LDA - HDP - DTM - DIM
() sinat_26917383 (http://blog.csdn.net/sinat_26917383) 2017年04月09日 11:23 口5561
漏气!重磅改革!吴恩达说:女儿识字后就教她学Python! Python的火爆最近越来越挡不住了,连身边多年工作经验的朋友都开始学Python了!他是这么说的
nttp://www.baidu.com/cb.php?c=lgF_pyfqnHmknjnvPjc0IZ0qnfK9ujYzP1ndPWb10Aw- Hc3rHnYnHb0TAq15HfLPWRznjb0T1dbrHw9mvm3ujDzn1Nbm1N90AwY5HDdnHnYnjcYPHm0lgF_5y9YIZ0IQzq- ZR8mLPbUB48ugfEIAqspynETZ-YpAq8nWqdIAdxTvqdThP-
yF_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqnHRLPjnvnfKEpyfqnHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqnHn3PW0) Gensim入门教程 (http://blog.csdn.net/Together_CZ/article/details/72888505)
转自:http://www.cnblogs.com/lioveai/p/gensim_tutorial.html What is Gensim? Gensim是一款开源的第三方Py
Together_CZ (http://blog.csdn.net/Together_CZ) 2017年06月06日 22:08 \$\text{Q348}\$
【gensim中文教程】开始使用gensim (http://blog.csdn.net/DuinoDu/article/details/766186
原文链接介绍了基本概念,以及理解和使用gensim的基本元素,并提供了一个简单的例子。
S DuinoDu (http://blog.csdn.net/DuinoDu) 2017年08月03日 14:08 □ 2417
gensim 教程 -Part1 (http://blog.csdn.net/Thinking_boy1992/article/details/53285592)
土中郊汉中 Consist (中区 No. 4) 中国 大大国的保护机由中沿马及孙大王,上了游江口土,"仁仁, No. 1 (1) (1)

ning >>> Ingging basicConfi





一学就会的 WordPress 实战课

认真学完这个系列课程之后,会深入了解 WordPress 的使用和开发,并掌握基本的 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=IgF_pyfqnHmknjTYPHf0IZ0qnfK9ujYzP1f4Pjnd0Aw-

5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YduH6znADsPjczn1w9nhwW0AwY5HDdnHnYnjcYPWD0IqF 5y9YIZ0IOzqMpqwBUvqoOhP8OvIGIAPCmqfEmvq Iyd8Q1N9nHmvnj7hnHPWnjFhPAD1Pyn4uW99ujqdIAdxTvqdThP-5 HDznHN9mhkEusKzujYk0AFV5H00TZcqn0KdpyfqnHRLPjnvnfKEpyfqnHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqn1Dkn1f)

gensim使用方法以及例子 (http://blog.csdn.net/u014595019/article/details/52218249)

gensim是一个python的自然语言处理库,能够将文档根据TF-IDF, LDA, LSI 等模型转化成向量模式,以便进行进一步的处 理。此外,gensim还实现了word2vec功能,能够将单词转...

🔊 u014595019 (http://blog.csdn.net/u014595019) 2016年08月16日 10:58 🔲 12397

 \triangle 内容坐报

TÔP 返回顶部

gensim做主題模型 (http://blog.csdn.net/luzhonghe1991/article/details/52196026)

作为python的一个库,gensim给了文本主题模型足够的方便,像他自己的介绍一样,topic modelling for humans 具体的tutori al可以参看他的官方...

【動 luzhonahe1991 (http://blog.csdn.net/luzhonahe1991) 2016年08月12日 22:55 □ 914

基于gensim的Doc2Vec简析 (http://blog.csdn.net/lenbow/article/details/52120230)

摘要:本文主要描述了一种文章向量(doc2vec)表示及其训练的相关内容,并列出相关例子。两位大牛Quoc Le 和 Tomas Mikolov (搞出Word2vec的家伙)在2014年的《Distr...

GENSIM 使用笔记2 — 主題模型和相似性查询 (http://blog.csdn.net/MebiuW/article/details/5...

GENSIM 使用笔记1 — 语料和向量空间 GENSIM 使用笔记2 — 主题模型和相似性查询 在上一个笔记当中,使用gensim针对 中文预料创建了字典和语料库,在这一章节中,主要讲下如何创建相..

MebiuW (http://blog.csdn.net/MebiuW) 2016年12月25日 17:04 □1977

gensim做主題模型 (http://blog.csdn.net/whzhcahzxh/article/details/17528261)

作为python的一个库,gensim给了文本主题模型足够的方便,像他自己的介绍一样,topic modelling for humans 具体的tutori al可以参看他的官方网页, 当然是全英...

Nhzhcahzxh (http://blog.csdn.net/whzhcahzxh) 2013年12月24日 15:28 山17213

Gensim实战(一) (http://blog.csdn.net/u013776640/article/details/42347983)

作为自然语言处理爱好者,大家都应该听说过或使用过大名鼎鼎的Gensim吧,这个一款具备多种功能的神器,为了深入了解 该工具的使用方法,本人将使用该工具进行一系列实战。 该系列博客共分为以下...

■ u013776640 (http://blog.csdn.net/u013776640) 2015年01月02日 23:24 単7317

【python gensim使用】word2vec词向量处理中文语料 (http://blog.csdn.net/churximi/articl...

word2vec介绍word2vec官网:https://code.google.com/p/word2vec/word2vec是google的一个开源工具,能够根据输入的词

 \triangle 内容坐报 ß

∆ 28 □

的集合计算出词与词之间的
Churximi (http://blog.csdn.net/churximi) 2016年05月21日 20:57 □□25106
Python版的Word2Vec gensim 学习手札 中文词语相似性度量 V1.1 (http://blog.csdn.net/M
前言相关内容链接:第一节:Google Word2vec 学习手札 昨天好不容易试用了一下Google自己提供的Word2Vector的源代 码,花了好长时间训练数据,结果发现似乎Python并不能…
■ MebiuW (http://blog.csdn.net/MebiuW) 2016年08月24日 20:10 □ □ 9398
Gensim做中文主題模型 (LDA) (http://blog.csdn.net/kunlong0909/article/details/48029111)
中文语料来自http://www.sogou.com/labs/dl/c.html 的精简版(tar.gz格式) 24M jerry@hq:/u01/jerry/Reduced\$ is C0000
Python与自然语言处理(二)基于Gensim的Word2Vec (http://blog.csdn.net/monkey13149
Java调用NLPIC实现分词和标注工作,Python调用Word2Vec实现词向量相似度计算
monkey131499 (http://blog.csdn.net/monkey131499) 2016年04月12日 10:13 □8524
如何安装python gensim (http://blog.csdn.net/Star_Bob/article/details/47835217)
本文前提是已经安装了python 的scipy和numpy模块,针对eclipse 编辑器的使用首先,下载python gensim包进入下载页面: http://download.csdn.net/d
Star_Bob (http://blog.csdn.net/Star_Bob) 2015年08月21日 15:25 □3754
Delphi7高级应用开发随书源码 (http://download.csdn.net/download/chenx
ZP
/http://download / 2003年04月30日 00:00 676KB 下载
2000 104 300 00.00 00.00
主题模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现lda模型。gensim简介Gensim是一个相当
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现da模型。gensim简介Gensim是一个相当专
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现lda模型。gensim简介Gensim是一个相当专 pipisorry (http://blog.csdn.net/pipisorry) 2015年06月10日 22:27 □8194
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现Ida模型。gensim简介Gensim是一个相当专
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现da模型。gensim简介Gensim是一个相当 专 pipisorry (http://blog.csdn.net/pipisorry) 2015年06月10日 22:27 口8194 python中用gensim做wiki的中文数据word2vector处理 (http://blog.csdn.net/lixintong1992/
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现da模型。gensim简介Gensim是一个相当 专 pipisorry (http://blog.csdn.net/pipisorry) 2015年06月10日 22:27 口8194 python中用gensim做wiki的中文数据word2vector处理 (http://blog.csdn.net/lixintong1992/
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现lda模型。gensim简介Gensim是一个相当 专 pipisorry (http://blog.csdn.net/pipisorry) 2015年06月10日 22:27 口8194 python中用gensim做wiki的中文数据word2vector处理 (http://blog.csdn.net/lixintong1992/ gensim 用conda install gensim 与pip install gensim 安装是不同的 提示C编译器会更快,装了 MinGW
主題模型TopicModel:通过gensim实现LDA (http://blog.csdn.net/pipisorry/article/details/4 http://blog.csdn.net/pipisorry/article/details/46447561使用python gensim轻松实现lda模型。gensim简介Gensim是一个相当专 pipisorry (http://blog.csdn.net/pipisorry) 2015年06月10日 22:27 □8194 python中用gensim做wiki的中文数据word2vector处理 (http://blog.csdn.net/lixintong1992/ gensim 用conda install gensim 与pip install gensim 安装是不同的 提示C编译器会更快,装了 MinGW