用深度学习获取文本语义：词向量应用于自然语言处理

1 新智元推荐

来源：OReillyData 授权转载

作者：Lior Shkiller

新智元为COO和执行总编提供最高超百万的年薪激励；为骨干员工提供最完整的培训体系、 高于业界平均水平的工资和奖金。

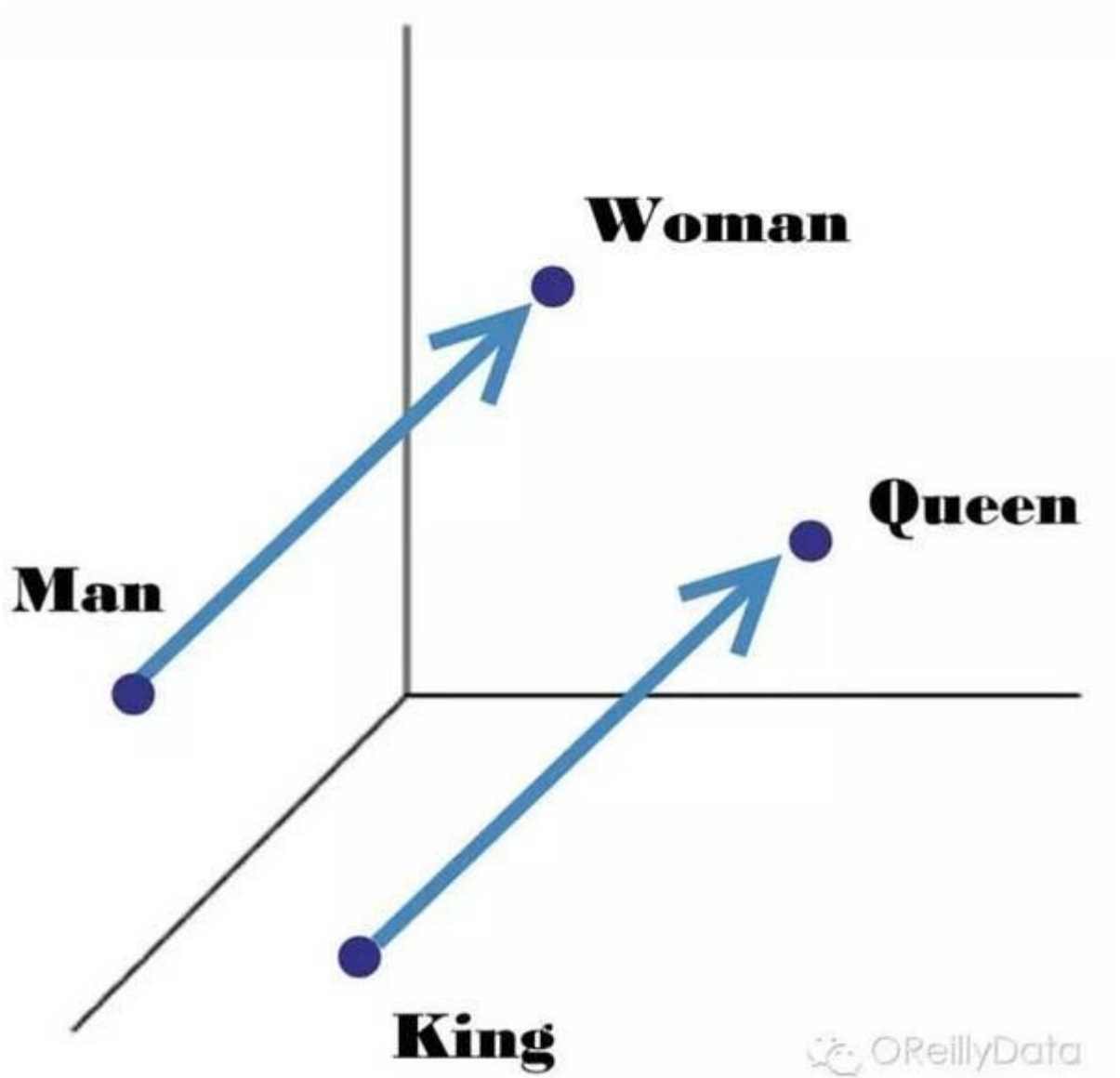
加盟新智元，与人工智能 (http://www.kejilie.com/channel/rengongzhineng.html)业界领袖携手改变世界。

【新智元导读】 词向量是一种把词处理成向量的技术，并且保证向量间的相对相似度和语义相似度是相关的。这个技术是在无监督学习方面最成功的应用之一。本文作者作为机器学习实践者，在文中介绍了如何编写一个神经网络模型来计算词间的关系并提高效率。结果表明，词向量确实能找到词汇之间的语义关系，还可以应用于更多领域。

词向量是一种把词处理成向量的技术，并且保证向量间的相对相似度和语义相似度是相关的。这个技术是在无监督学习方面最成功的应用之一。传统上， 自然语言处理 (http://www.kejilie.com/channel/ziranyuyanchuli.html)（NLP） 系统把词编码成字符串。这种方式是随意确定的，且对于获取词之间可能存在的关系并没有提供有用的信息。词向量是NLP领域的一个替代方案。它把词或短语映射成实数向量，把特征从词汇表大小的高维度空间降低到一个相对低的维度空间。

例如，让我们看看四个词：“woman”（女人）、“man”（男人）、“queen”（女王）和“king”（国王）。我们把它们都向量化，再使用简单的代数运算来发现它们之间的语义相似度。计算向量间的相似度可以采用诸如余弦相似度的方法。当我们把词“woman”的向量减去词“man”后，这个差值的余弦相似度应该和词“queen”的向量减去“king”的向量的差值比较接近（参见图1）。

W(“woman”)?W(“man”) ? W(“queen”)?W(“king”)



提交建议(http://cn.mikecrm.com/amttlxP)

图1 性别的向量。来源：Lior Shkiller

有很多不同的模型可以被用来把词转换成实数性的向量，包括 隐含语义分析（LSA）和 隐含狄利克雷分布（LDA）。这些模型背后的思路是：相关的词汇一般都会在相同的文档里同时出现。例如，backpack（背包）、school（学校）、notebook（笔记本）和teacher（教师）一般都会一起出现。而school（学校）、tiger（老虎）、apple（苹果）和basketball（篮球）一般都不会持续同时出现。基于这个相关的词会在相关的文档里出现的基本假设，为了把词转化为向量，LSA会构建一个矩阵。矩阵的行是（语料库或数据里）所有出现过的词，而列则是对应于文档里的一个段落。LSA使用 奇异值分解（SVD）的方法，在保存列之间相似性的同时降低矩阵的行数。不过这些模型的主要问题是：在数据量非常大的时候，计算量也非常得大。

为了避免计算和存储大量的数据，我们试图创造一个神经网络模型来计算词间的关系，并提高效率。

Word2Vec

目前最流行的词向量模型是由 Mikolov等人 在2013年提出的 word2vec 。这个模型的效果很好，且计算效率有了很大的提升。Mikolov等提出的负采样方法是一个更有效的产生词向量的方法。更多的信息可以在 这里找到。

这一模型可以使用下述两种架构的任一种来生成词的分布： 连续词袋（CBOW）和 连续跳跃元语法（skip-gram）。

下面让我们分别来看看这两种架构。

CBOW模型

在CBOW架构里，模型根据目标词的上下文来预测目标词。因此，Mikolov等使用了目标词 *w* 的前 *n* 个词和后 *n* 个词。

一个序列的词等同于一个物品集。因此，就可以把“词”理解为“物品”。对于“物品”我们可以使用 推荐系统 以及 协同过滤 里的方法。CBOW模型的训练速度是跳跃元语法模型的七倍，而且预测准确性也稍好（参见图2）。

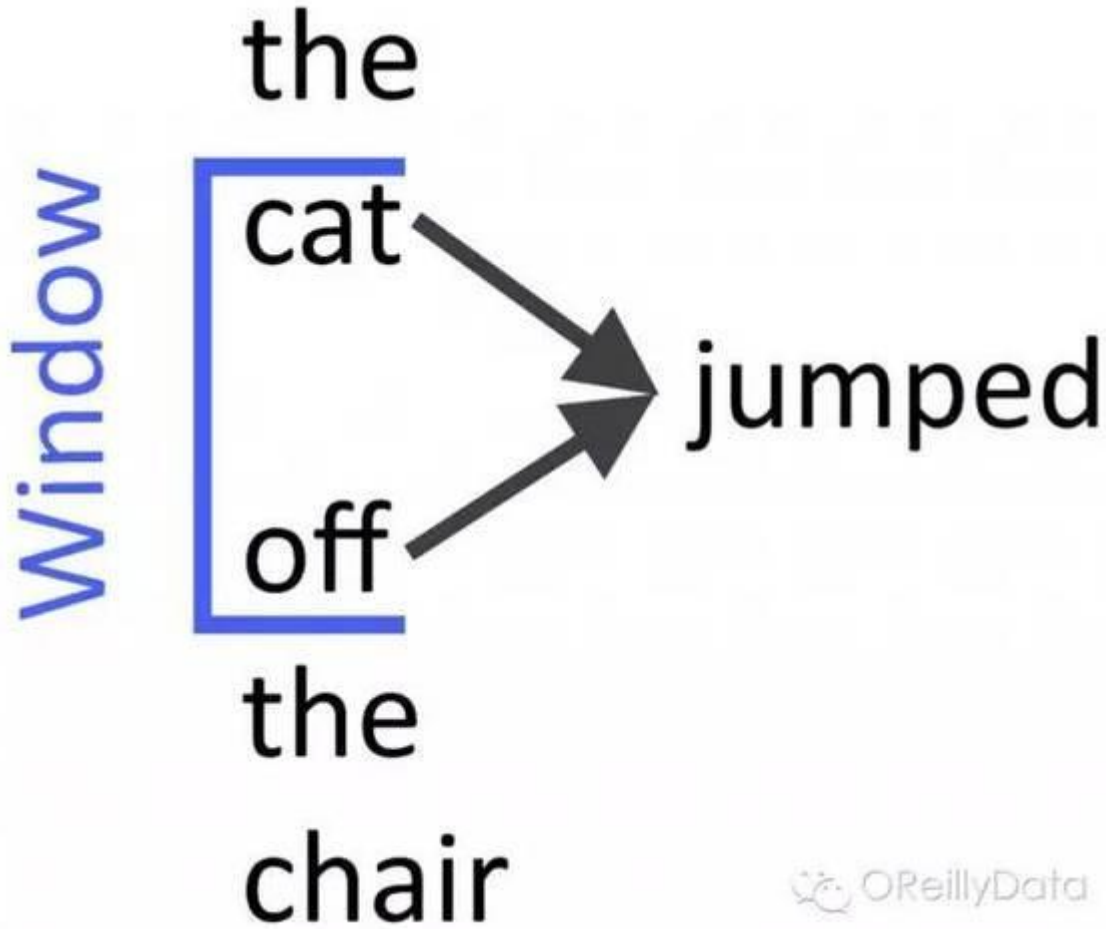


图2 基于上下文来预测词。来源：Lior Shkiller

连续跳跃元语法模型

与使用目标词的上下文的方法不同，连续跳跃元语法模型是使用目标词去预测它的前后词（参见图3）。据Mikolov等的论文，在训练数据量比较小的时候，跳跃元语法模型比较好，且对于罕见的词和短语的处理较好。

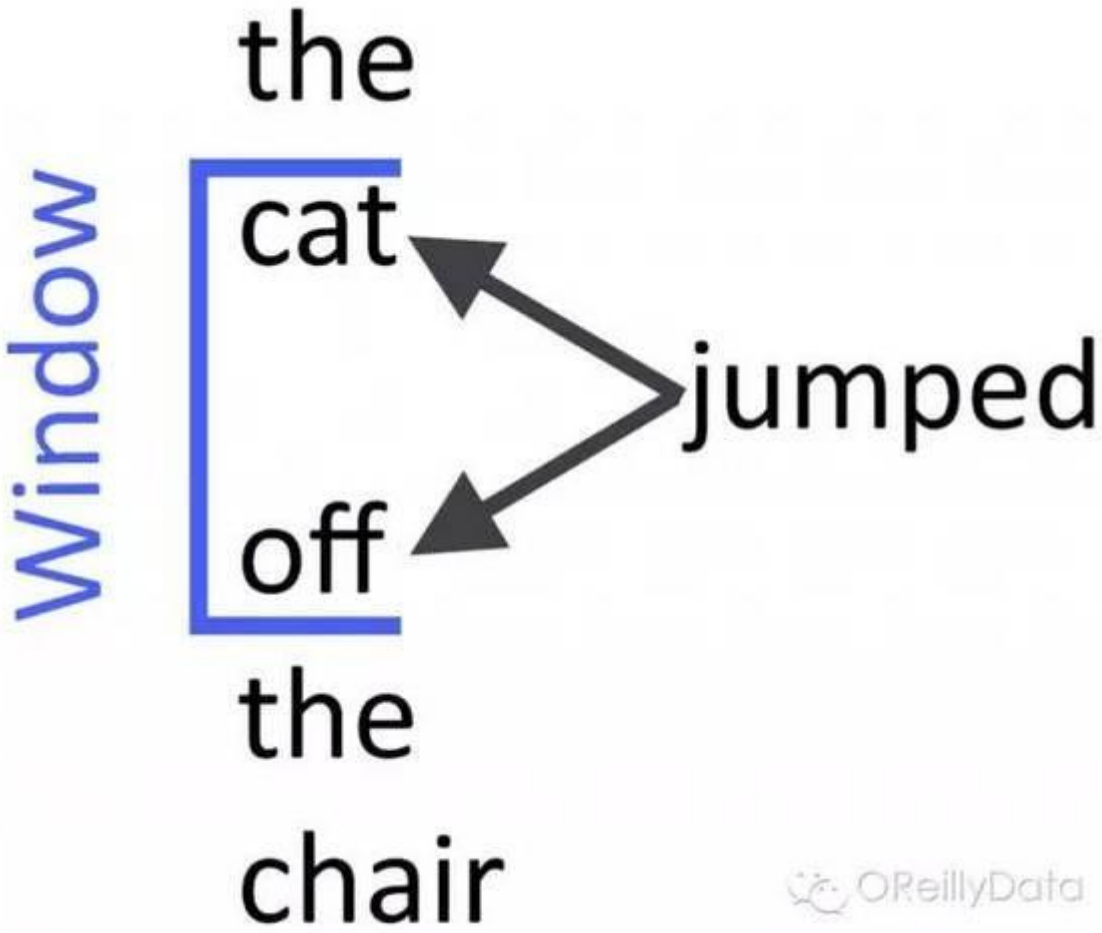


图3 用给定的词来预测上下文。来源：Lior Shkiller

代码

（你可以在这个 GitHub库 里找到下面例子的代码）

这个模型（word2vec）的一大好处就是，它可以用于很多种语言。

我们所要做的就是下载一个所要处理的语言的大数据集。

从维基百科上找一个大数据集

我们可以从维基百科里面找到很多语言的数据。用下面的步骤就可以获得一个大数据集。

- 找到你想处理的语言的ISO 639代码： ISO 639代码 的列表
- 登录https://dumps.wikimedia.org/wiki/latest/（译者注：此链接已失效）
- *下载 wiki-latest-pages-articles.xml.bz2*

接着，为了让后续的事情变简单，我们会安装 gensim 。它是一个实现了word2vec的Python库。

pip install ﷿upgrade gensim

我们需要用维基百科的下载文件来创建语料库，以用于后续的word2vec模型的训练。下面这段代码的输出就是一个“wiki..text”的文件。其中包括了维基百科的所有文章的所有词汇，并按照语言分开。

from gensim.corpora import WikiCorpus

```
language_code = “he”

inp = language_code+”wiki-latest-pages-articles.xml.bz2″

outp = “wiki.{}.text”.format(language_code)

i = 0

print(“Starting to create wiki corpus”)

output = open(outp, ‘w’)

space = ” ”

wiki = WikiCorpus(inp, lemmatize=False, dictionary={})

for text in wiki.get_texts():

article = space.join([t.decode(“utf-8”) for t in text])

output.write(article + “n”)

i = i + 1
```

```
if (i % 1000 == 0):

print(“Saved ” + str(i) + ” articles”)

output.close()

print(“Finished  Saved ” + str(i) + ” articles”)
```

训练模型

参数的说明如下：

- size：向量的维度
  - 大的size值会要求更多的训练数据，但能带来更准确的模型
- window：在一个句子内，目标词与预测词之间的最大距离
- min_count：忽略所有总词频低于这个值的词。

```
import multiprocessing

from gensim.models import Word2Vec
```

```
from gensim.models.word2vec import LineSentence

language_code = “he”

inp = “wiki.{}.text”.format(language_code)

out_model = “wiki.{}.word2vec.model”.format(language_code)

size = 100

window = 5

min_count = 5

start = time.time()

model = Word2Vec(LineSentence(inp), sg = 0, # 0=CBOW , 1= SkipGram

size=size, window=window, min_count=min_count, workers=multiprocessing.cpu_count())

# trim unneeded model memory = use (much) less RAM

model.init_sims(replace=True)
```

print(time.time()-start)

model.save(out_model)

整个word2vec训练过程用了18分钟。

fastText库

Facebook的人工智能研究（FAIR）实验室最近发布了 fastText库 。它是基于 Bojanowski等 的论文《Enriching Word Vectors with Subword Information 》所开发的模型。与word2vec不同，fastText把词表示成一个n元的字母袋。每个向量代表字符袋里的一个n元字母，而一个词则是这些向量的和。

使用的新库很简单。安装命令：

pip install fasttext

训练模型的命令：

start = time.time()

language_code = “he”

inp = “wiki.{}.text”.format(language_code)

output = “wiki.{}.fasttext.model”.format(language_code)

model = fasttext.cbow(inp,output)

print(time.time()-start)

整个fastText模型训练用了13分钟。

评估向量：类比性

下面让我们用之前的那个例子来评估这两个模型的准确度。

W(“woman”) ? W(“man”)+ W(“queen”)? W(“king”)

下面的代码首先计算正负词的加权平均值。

随后，代码计算了所有的测试词汇的向量与加权平均的点乘积。

我们的评估例子里，测试词汇是整个词汇表。代码的最后是打印出和正词与负词的加权平均值的余弦相似度最高的词。

import numpy as np

```
from gensim.matutils import unitvec

def test(model,positive,negative,test_words):

mean = []

for pos_word in positive:

mean.append(1.0 * np.array(model[pos_word]))

for neg_word in negative:

mean.append(-1.0 * np.array(model[neg_word]))

# compute the weighted average of all words

mean = unitvec(np.array(mean).mean(axis=0))

scores = {}

for word in test_words:

if word not in positive + negative:
```

```
test_word = unitvec(np.array(model[word]))

# Cosine Similarity

scores[word] = np.dot(test_word, mean)

print(sorted(scores, key=scores.get, reverse=True)[:1])
```

接着，用我们最初的那个例子来做测试。

用fastText和gensim的word2vec模型来预测：

```
positive_words = [“queen”,”man”]

negative_words = [“king”]

# Test Word2vec

print(“Testing Word2vec”)

model = word2vec.getModel()

test(model,positive_words,negative_words,model.vocab)
```

```
# Test Fasttext

print("Testing Fasttext")

model = fasttxt.getModel()

test(model,positive_words,negative_words,model.words)
```

结果

```
Testing Word2vec

['woman']

Testing Fasttext

['woman']
```

结果显示fastText和gensim的word2vec都能正确预测。

W("woman") ? W("man")+ W("queen")? W("king")

可见，词向量确实能找到词汇之间的语义关系。

我们这里所介绍的模型的基本思路可以被运用到很多的应用场景。如 预测商业机构需要的下一个应用 、 做情感分析 、 替换生物序列 、做 语义图片搜索 等。

【作者介绍】Lior Shkiller 是Deep Solution的联合创始人。作为一个机器学习的实践者，他积极热忱地投身于人工智能和认知科学。Lior拥有以色列特拉维夫大学的计算机科学与心理学学位，并有超过10年的软件开发经验。Deep Solutions提供端到端的软件解决方案，其中包括为计算机视觉、自然语言处理、异常检测和推荐系统等应用所开发的创新的深度学习的新算法。

新智元招聘

职位 运营总监

职位年薪：36- 50万（工资+奖金）

工作地点：北京-海淀区

所属部门：运营部

汇报对象：COO

下属人数：2人

年龄要求：25 岁 至 35 岁

性别要求：不限

工作年限：3 年以上

语 言：英语6级（海外留学背景优先）

职位描述

1. 负责大型会展赞助商及参展商拓展、挖掘潜在客户等工作，人工智能及机器人产业方向

2. 擅长开拓市场，并与潜在客户建立良好的人际关系

3. 深度了解人工智能及机器人产业及相关市场状况，随时掌握市场动态

4. 主动协调部门之间项目合作，组织好跨部门间的合作，具备良好的影响力

5. 带领团队完成营业额目标，并监控管理项目状况

6. 负责公司平台运营方面的战略计划、合作计划的制定与实施

岗位要求

1. 大学本科以上学历，硕士优先，要求有较高英语沟通能力

2. 3年以上商务拓展经验，有团队管理经验，熟悉商务部门整体管理工作

3. 对传统全案公关、传统整合传播整体方案、策略性整体方案有深邃见解

4. 具有敏锐的市场洞察力和精确的客户分析能力、较强的团队统筹管理能力

5. 具备优秀的时间管理、抗压能力和多任务规划统筹执行能力

6. 有广泛的TMT领域人脉资源、有甲方市场部工作经验优先考虑

7. 有媒体广告部、市场部，top20公关公司市场拓展部经验者优先

新智元欢迎有志之士前来面试，更多招聘岗位请访问新智元公众号。

随意打赏


(http://www.kejilie.com/36dsj/article/jyMR3u.html)

RNN在自然语言处理中的应用及其PyTorch实现 (http://www.kejilie.com/36dsj/article/jyMR3u.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 5天前

作者：廖星宇对于人类而言，以前见过的事物会在脑海里面留下记忆，虽然随后记忆会慢慢消失，但是每当经过提醒，人们往往能够重拾记忆。在神经网络的研究中，让模型充满记忆力的研究很早便开始了，Saratha Sathasivam 于1982 年提出了霍普菲尔德网络，但是由于它实现困难，在提出的时…


(http://www.kejilie.com/36dsj/article/iMjiE3.html)

大数据早报：海量大数据重度孵化器获A+轮融资 阿里自然语言处理技术获突破（11.30） (http://www.kejilie.com/36dsj/article/iMjiE3.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 5天前

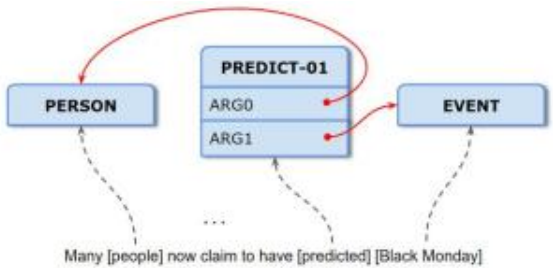
数据早知道，上36dsj看早报！来源36大数据，作者：奥兰多『融资』海量大数据重度孵化器获A+轮融资，估值超1亿美元11月29日消息，海量大数据重度孵化器宣布公司已于2017年10月获得了广州众上集团的A+轮投资。据悉，这是海量大数据重度孵化器成立以来，一年内获得的第二轮融资。完成..

(http://www.kejilie.com/iyiou/article/eqiUje.html)

病历智能分析系统：挖掘自然语言处理技术在医疗大数据中的价值 (http://www.kejilie.com/iyiou/article/eqiUje.html)

亿欧网 (http://www.kejilie.com/iyiou/index.html) • 7天前

病历 作为医院的宝贵财富，里面蕴含了大量的专业知识，但是由于受到技术的限制，长期得不到有效利用。我院作为一所大型三甲综合医院，学科实力雄厚，对临床科研的要求也非常高。但是医生还停留在去病案室借阅病历，手工摘抄收集科研数据的阶段，效率十分低下。如何利用最新的 人工智…



(http://www.kejilie.com/leiphone/article/QR7nUv.html)

Google发布自然语言处理解析器SLING，免除模块化分析级联效应产生的缺陷 (http://www.kejilie.com/leiphone/article/QR7nUv.html)

雷锋网 (http://www.kejilie.com/leiphone/index.html) • 19天前

雷锋网 AI科技评论消息，日前，Google发布自然语言框架语义解析器SLING，它能以语义框架图（semantic frame graph）的形式，将自然语言文本直接解析为文本语义表示。这一系统避免了级联效应，另外还减少了不必要的计算开销。详细消息雷锋网 AI科技评论编译整理如下：直到最近，大多数…

 (http://www.kejilie.com/36dsj/article/nUf2im.html)

最全面的百度NLP自然语言处理技术解析 (http://www.kejilie.com/36dsj/article/nUf2im.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 21天前

作者： 田宁宁在AI时代，我们希望计算机能够拥有视觉、听觉、行动以及语言的智能，而相对于听和看以及行动，语言是我们人类区别于其他动物的最重要特征之一。语言是我们思维的载体，也因此我们对于语言的理解和处理，变得尤为重要。而在计算机领域，自然语言处理（NLP, Natural…

 (http://www.kejilie.com/36dsj/article/N7BBfi.html)

号称世界最快句法分析器，Python高级自然语言处理库spaCy (http://www.kejilie.com/36dsj/article/N7BBfi.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 22天前

spaCy是Python和Cython中的高级自然语言处理库，它建立在最新的研究基础之上，从一开始就设计用于实际产品。spaCy带有预先训练的统计模型和单词向量，目前支持20多种语言的标记。它具有世界上速度最快的句法分析器，用于标签的卷积神经网络模型，解析和命名实体识别以及与深度学习…


(http://www.kejilie.com/36dsj/article/jeaqi2.html)

外行也能看懂的科普：这就叫自然语言处理 (http://www.kejilie.com/36dsj/article/jeaqi2.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 1月前

作者：武汉飓拓科技一、什么是自然语言处理简单地说，自然语言处理（NaturalLanguage Processing，简称NLP）就是用计算机来处理、理解以及运用人类语言(如中文、英文等)，它属于人工智能的一个分支，是计算机科学与语言学的交叉学科，又常被称为计算语言学。由于自然语言是人类区别…


(http://www.kejilie.com/36dsj/article/eIVVvy.html)

自然语言处理技术，将会使机器从更人性化的视角来解决问题 (http://www.kejilie.com/36dsj/article/eIVVvy.html)

36大数据 (http://www.kejilie.com/36dsj/index.html) • 1月前

作者： chiming在解决和数学或物理相关的问题时，技术能够发挥相当大的作用。但在解决涉及到以人为中心的问题时，技术的可发挥余地就变得很小。在Ultimate Software的高级策略总监Armen Berjikly看来，自然语言处理（NLP）的进步会帮助技术更好的识别出人类的情感与同情心，从而来

 (http://www.kejilie.com/leiphone/article/6vuEby.html)

想要用好自然语言处理技术，先要克服这些困难！ **(http://www.kejilie.com/leiphone/article/6vuEby.html)**

雷锋网 (http://www.kejilie.com/leiphone/index.html) • 1月前

雷锋网按 ：10月11日-14日在杭州举办的的云栖大会上，马云公布达摩院的研究领域包括：量子计算、机器学习、自然语言处理、基础算法、等前沿技术再次掀起了前沿科技讨论的浪潮。 人工智能已经是大部分普通人都耳熟能详的词汇，而人们对自然语言处理技术的了解程度却大部分还停留在表面…

 (http://www.kejilie.com/iyiou/article/3qUBbe.html)

自然语言处理如何助力人机共鸣 **(http://www.kejilie.com/iyiou/article/3qUBbe.html)**

亿欧网 (http://www.kejilie.com/iyiou/index.html) • 1月前

本文来自venturebeat，作者刘敏；由亿欧编译。 在当前飞速发展的创新步伐中， 科技 似乎正在积极地解决人类最紧迫的难题。 在某些方面，我们取得了很大的进步。在可再生能源、疾病预防和灾后重建等领域作出了重大突破。但是，当涉及到解决如员工多样性、无意识偏见、员工和客户满意度…

评论