**[python+spaCy 进行简易自然语言处理](http://blog.csdn.net/u201017971/article/details/75675976)**

0．简介：

自然语言处理（NLP）是人工智能领域最重要的部分之一。它在许多智能应用中担任了关键的角色，例如聊天机器人、正文提取、多语翻译以及观点识别等应用。业界 NLP 相关的公司都意识到了，处理非结构文本数据时，不仅要看正确率，还需要注意是否能快速得到想要的结果。

NLP 是一个很宽泛的领域，它包括了文本分类、实体识别、机器翻译、问答系统、概念识别等子领域。在我最近的一篇[文章](https://link.juejin.im/?target=https%3A%2F%2Fwww.analyticsvidhya.com%2Fblog%2F2017%2F01%2Fultimate-guide-to-understand-implement-natural-language-processing-codes-in-python%2F)中，我探讨了许多用于实现 NLP 的工具与组件。在那篇文章中，我更多的是在描述[NLTK](https://link.juejin.im/?target=http%3A%2F%2Fwww.nltk.org%2F)（Natural Language Toolkit）这个伟大的库。

在这篇文章中，我会将 spaCy-这个现在最强大、最先进的 NLP python 库分享给你们。

1. spaCy简介及安装方法

1.1 简介

spaCy 由 cython（Python 的 C 语言拓展，旨在让 python 程序达到如同 C 程序一样的性能）编写，因此它的运行效率非常高。spaCy 提供了一系列简洁的 API 方便用户使用，并基于已经训练好的机器学习与深度学习模型实现底层。

1.2 安装

spaCy 及其数据和模型可以通过 pip 和安装工具轻松地完成安装。使用下面的命令在电脑中安装 spaCy：

sudo pip install spacy

如果你使用的是 Python3，请用 “pip3” 代替 “pip”。

或者你也可以在 [这儿](https://link.juejin.im/?target=https%3A%2F%2Fpypi.python.org%2Fpypi%2Fspacy) 下载源码，解压后运行下面的命令安装：

python setup.py install

在安装好 spacy 之后，请运行下面的命令以下载所有的数据集和模型：

python -m spacy.en.download all

一切就绪，现在你可以自由探索、使用 spacy 了。

**2. spaCy 的管道（Pipeline）与属性（Properties）**

spaCy 的使用，以及其各种属性，是通过创建管道实现的。在加载模型的时候，spaCy 会将管道创建好。在 spaCy 包中，提供了各种各样的[模块](https://link.juejin.im?target=https%3A%2F%2Fgithub.com%2Fexplosion%2Fspacy-models%2F" \t "_blank)，这些模块中包含了各种关于词汇、训练向量、语法和实体等用于语言处理的信息。

下面，我们会加载默认的模块（english-core-web 模块）。

import spacy

nlp = spacy.load(“en”)

“nlp” 对象用于创建 document、获得语言学标注（linguistic annotation）及其它的 nlp 属性。首先我们要创建一个document，将文本数据加载进管道中。我使用了来自猫途鹰网的旅店评论数据。这个数据文件可以在[这儿](https://link.juejin.im?target=https%3A%2F%2Fs3-ap-south-1.amazonaws.com%2Fav-blog-media%2Fwp-content%2Fuploads%2F2017%2F04%2F04080929%2FTripadvisor_hotelreviews_Shivambansal.txt" \t "_blank)下载。

document = unicode(open(filename).read().decode('utf8'))

document = nlp(document)

这个 document 现在是 spacy.english 模型的一个 class，并关联上了许多的属性。可以使用下面的命令列出所有 document（或 token）的属性：

dir(document)

>> [ 'doc', 'ents', … 'mem']

它会输出 document 中各种各样的属性，例如：token、token 的 index、词性标注、实体、向量、情感、单词等。下面让我们会对其中的一些属性进行一番探究。

**2.1 Tokenization**

spaCy 的 document 可以在 tokenized 过程中被分割成单句，这些单句还可以进一步分割成单词。你可以通过遍历文档来读取这些单词：

作者：lsvih  
链接：https://juejin.im/post/5971a4b9f265da6c42353332  
来源：掘金  
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。