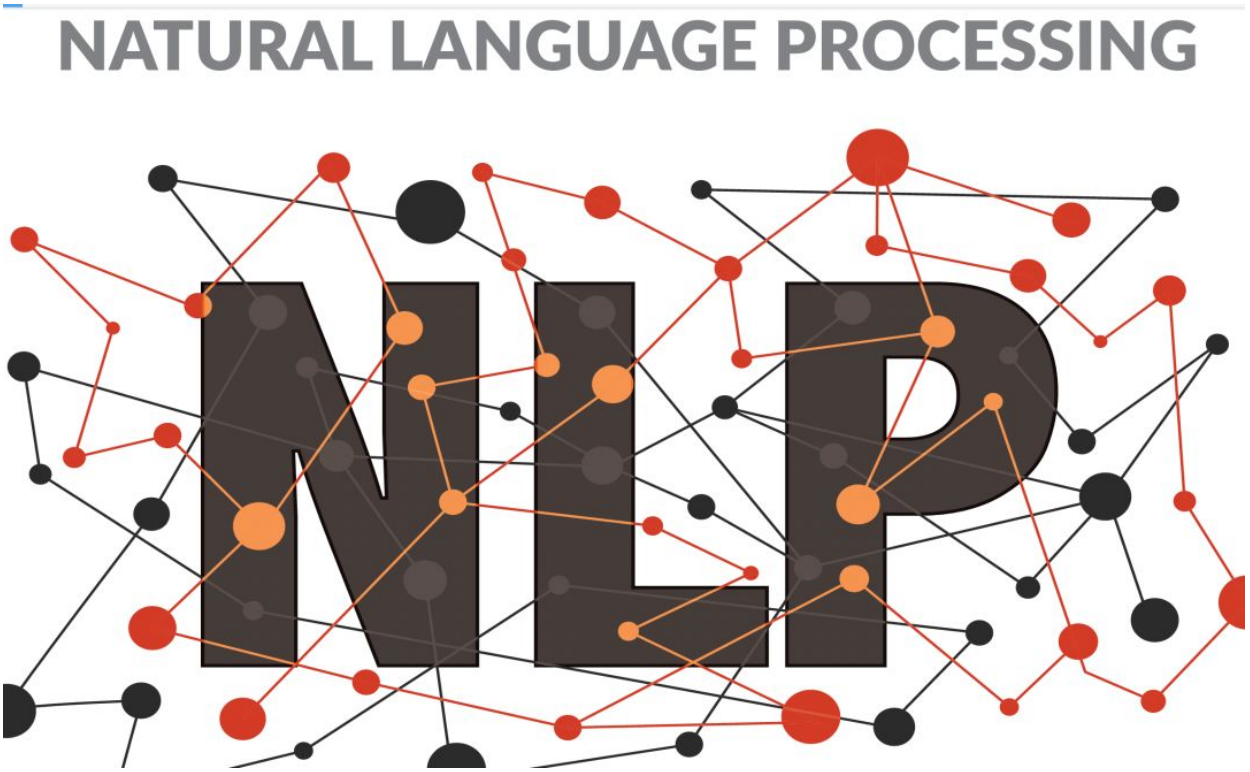
自然语言处理NLP快速入门



计算机非常擅长处理标准化和结构化的数据，如数据库表和财务记录。他们能够比我们人类更快地处理这些数据。但我们人类不使用“结构化数据”进行交流，也不会说二进制语言！我们用文字进行交流，这是一种非结构化数据。

不幸的是，计算机很难处理非结构化数据，因为没有标准化的技术来处理它。当我们使用c、java或python之类的语言对计算机进行编程时，我们实际上是给计算机一组它应该操作的规则。对于非结构化数据，这些规则是非常抽象和具有挑战性的具体定义。

互联网上有很多非结构化的自然语言，有时甚至连谷歌都不知道你在搜索什么！

人与计算机对语言的理解

人类写东西已经有几千年了。在这段时间里，我们的大脑在理解自然语言方面获得了大量的经验。当我们在一张纸上或互联网上的博客上读到一些东西时，我们就会明白它在现实世界中的真正含义。我们感受到了阅读这些东西所引发的情感，我们经常想象现实生活中那东西会是什么样子。

自然语言处理（NLP）是人工智能的一个子领域，致力于使计算机能够理解和处理人类语言，使计算机更接近于人类对语言的理解。计算机对自然语言的直观理解还不如人类，他们不能真正理解语言到底想说什么。简而言之，计算机不能在字里行间阅读。

尽管如此，机器学习（ML）的最新进展使计算机能够用自然语言做很多有用的事情！深度学习使我们能够编写程序来执行诸如语言翻译、语义理解和文本摘要等工作。所

有这些都增加了现实世界的价值，使得你可以轻松地理解和执行大型文本块上的计算，而无需手工操作。

让我们从一个关于NLP如何在概念上工作的快速入门开始。之后，我们将深入研究一些python代码，这样你就可以自己开始使用NLP了！

NLP难的真正原因

阅读和理解语言的过程比乍一看要复杂得多。要真正理解一段文字在现实世界中意味着什么，有很多事情要做。例如，你认为下面这段文字意味着什么？

> “Steph Curry was on fire last nice. He totallydestroyed the other team”

对一个人来说，这句话的意思很明显。我们知道 Steph Curry 是一名篮球运动员，即使你不知道，我们也知道他在某种球队，可能是一支运动队。当我们看到“着火”和“毁灭”时，我们知道这意味着Steph Curry昨晚踢得很好，击败了另一支球队。

计算机往往把事情看得太过字面意思。从字面上看，我们会看到“Steph Curry”，并根据大写假设它是一个人，一个地方，或其他重要的东西。但后来我们看到Steph Curry“着火了”…电脑可能会告诉你昨天有人把Steph Curry点上了火！…哎呀。在那之后，电脑可能会说，curry已经摧毁了另一支球队…它们不再存在…伟大的…

Steph Curry真的着火了！

但并不是机器所做的一切都是残酷的，感谢机器学习，我们实际上可以做一些非常聪明的事情来快速地从自然语言中提取和理解信息！让我们看看如何在几行代码中使用几个简单的python库来实现这一点。

使用Python代码解决NLP问题

为了了解NLP是如何工作的，我们将使用Wikipedia中的以下文本作为我们的运行示例：

> Amazon.com, Inc., doing business as Amazon, is an Americanelectronic commerce and cloud computing company based in Seattle, Washington,that was founded by Jeff Bezos on July 5, 1994. The tech giant is the largestInternet retailer in the world as measured by revenue and market capitalization,and second largest after Alibaba Group in terms of total sales. The amazon.comwebsite started as an online bookstore and later diversified to sell videodownloads/streaming, MP3 downloads/streaming, audiobook downloads/streaming,software, video games, electronics, apparel, furniture, food, toys, andjewelry. The company also produces consumer electronics—Kindle e-readers,Fire tablets, Fire TV, and Echo

—and is the world’s largest provider of cloud infrastructure services (IaaS andPaaS). Amazon also sells certain low-end products under its in-house brandAmazonBasics.

几个需要的库

首先，我们将安装一些有用的python NLP库，这些库将帮助我们分析本文。

```
### Installing spaCy, general Python NLP lib

pip3 install spacy

### Downloading the English dictionary model for spaCy

python3 -m spacy download en_core_web_lg

### Installing textacy, basically a useful add-on to spaCy

pip3 install textacy
```

实体分析

现在所有的东西都安装好了，我们可以对文本进行快速的实体分析。实体分析将遍历文本并确定文本中所有重要的词或“实体”。当我们说“重要”时，我们真正指的是具有某种真实世界语义意义或意义的单词。

请查看下面的代码，它为我们进行了所有的实体分析：

```
# coding: utf-8

import spacy
```

```
### Load spaCy's English NLP model

nlp = spacy.load('en_core_web_lg')

### The text we want to examine

text = "Amazon.com, Inc., doing business as Amazon,
is anAmerican electronic commerce and cloud computing
company based in Seattle,Washington, that was founded
by Jeff Bezos on July 5, 1994. The tech giant isthe
largest Internet retailer in the world as measured by
revenue and marketcapitalization, and second largest
after Alibaba Group in terms of total sales.The amazon.
com website started as an online bookstore and later
diversified tosell video downloads/streaming, MP3
downloads/streaming, audiobookdownloads/streaming,
software, video games, electronics, apparel, furniture,
food, toys, and jewelry. The company also produces
consumer electronics-Kindle e-readers,Fire tablets,
Fire TV, and Echo-and is the world's largest provider
of cloud infrastructureservices (IaaS and PaaS).
Amazon also sells certain low-end products under
itsin-house brand AmazonBasics."

### Parse the text with spaCy
```

```
### Our 'document' variable now contains a parsed version
oftext.

document = nlp(text)

### print out all the named entities that were detected

for entity in document.ents:

    print(entity.text,entity.label_)
```

我们首先加载spaCy's learned ML模型，并初始化想要处理的文本。我们在文本上运行ML模型来提取实体。当运行taht代码时，你将得到以下输出：

```
Amazon.com, Inc. ORG
Amazon ORG
American NORP
Seattle GPE
Washington GPE
Jeff Bezos PERSON
July 5, 1994 DATE
second ORDINAL
Alibaba Group ORG
amazon.com ORG
Fire TV ORG
Echo -  LOC
PaaS ORG
Amazon ORG
AmazonBasics ORG
```

文本旁边的3个字母代码[1]是标签，表示我们正在查看的实体的类型。看来我们的模型干得不错！Jeff Bezos确实是一个人，日期是正确的，亚马逊是一个组织，西雅图和华盛顿都是地缘政治实体(即国家、城市、州等)。唯一棘手的问题是，Fire TV和Echo之类的东西实际上是产品，而不是组织。然而模型错过了亚马逊销售的其他产品“视频下载/流媒体、mp3下载/流媒体、有声读物下载/流媒体、软件、视频游戏、电子产品、服装、家具、食品、玩具和珠宝”，可能是因为它们在一个庞大的的列表中，因此看起来相对不重要。

总的来说，我们的模型已经完成了我们想要的。想象一下，我们有一个巨大的文档，里面满是几百页的文本，这个NLP模型可以快速地让你了解文档的内容以及文档中的关键实体是什么。

对实体进行操作

让我们尝试做一些更适用的事情。假设你有与上面相同的文本块，但出于隐私考虑，你希望自动删除所有人员和组织的名称。spaCy库有一个非常有用的清除函数，我们可以使用它来清除任何我们不想看到的实体类别。如下所示：

```
# coding: utf-8

import spacy

### Load spaCy's English NLP model
nlp = spacy.load('en_core_web_lg')

### The text we want to examine
text = "Amazon.com, Inc., doing business as Amazon,
is an American electronic commerce and cloud computing
company based in Seattle, Washington, that was founded
by Jeff Bezos on July 5, 1994. The tech giant is the
```

```
largest Internet retailer in the world as measured by
revenue and market capitalization, and second largest
after Alibaba Group in terms of total sales. The
amazon.com website started as an online bookstore and
later diversified to sell video downloads/streaming,
MP3 downloads/streaming, audiobook downloads/streaming,
 software, video games, electronics, apparel, furniture
 , food, toys, and jewelry. The company also produces
 consumer electronics-Kindle e-readers, Fire tablets,
   Fire TV, and Echo-and is the world's largest
   provider of cloud infrastructure services (IaaS and
   PaaS). Amazon also sells certain low-end products
   under its in-house brand AmazonBasics."

### Replace a specific entity with the word "PRIVATE"
def replace_entity_with_placeholder(token):
    if token.ent_iob != 0 and (token.ent_type_ == "PERSON" or
token.ent_type_ == "ORG"):
        return "[PRIVATE] "
      else:
        return token.string

### Loop through all the entities in a piece of text and apply
entity replacement
def scrub(text):
    doc = nlp(text)
    for ent in doc.ents:
```

```
        ent.merge()
    tokens = map(replace_entity_with_placeholder, doc)
    return "".join(tokens)

print(scrub(text))
```

```
[PRIVATE] , doing business as [PRIVATE] , is an American electronic
commerce and cloud computing company based in Seattle, Washington,
that was founded by [PRIVATE] on July 5, 1994. The tech giant is the
largest Internet retailer in the world as measured by revenue and
market capitalization, and second largest after [PRIVATE] in terms
of total sales. The [PRIVATE] website started as an online bookstore
and later diversified to sell video downloads/streaming, MP3
downloads/streaming, audiobook downloads/streaming, software, video
games, electronics, apparel, furniture, food, toys, and jewelry. The
company also produces consumer electronics - Kindle e-readers, Fire
tablets, [PRIVATE] , and Echo - and is the world's largest provider
of cloud infrastructure services (IaaS and [PRIVATE] ). [PRIVATE]
also sells certain low-end products under its in-house brand
[PRIVATE] .
```

效果很好！这实际上是一种非常强大的技术。人们总是在计算机上使用ctrl+f函数来查找和替换文档中的单词。但是使用NLP，我们可以找到和替换特定的实体，考虑到它们的语义意义，而不仅仅是它们的原始文本。

从文本中提取信息

我们之前安装的textacy库在spaCy的基础上实现了几种常见的NLP信息提取算法。它会让我们做一些比简单的开箱即用的事情更先进的事情。

它实现的算法之一是半结构化语句提取。这个算法从本质上分析了spaCy的NLP模型能够提取的一些信息，并在此基础上获取一些关于某些实体的更具体的信息！简而言之，我们可以提取关于我们选择的实体的某些“事实”。

让我们看看代码中是什么样子的。对于这一篇，我们将把华盛顿特区维基百科页面的全部摘要都拿出来。

```
# coding: utf-8

import spacy
import textacy.extract

### Load spaCy's English NLP model
nlp = spacy.load('en_core_web_lg')

### The text we want to examine
text = """Washington, D.C., formally the District of Columbia and commonly referred to as Washington or D.C., is the capital of the United States of America.[4] Founded after the American Revolution as the seat of government of the newly independent country, Washington was named after George Washington, first President of the United States and Founding Father.[5]
Washington is the principal city of the Washington metropolitan area, which has a population of 6,131,977.[6] As the seat of the United States federal government and several international organizations, the city is an important world political capital.[7] Washington is one of the most visited cities in the world, with more than 20 million annual tourists.[8][9]
The signing of the Residence Act on July 16, 1790, approved the creation of a capital district located along the Potomac River on the country's East Coast. The U.S. Constitution
```

provided for a federal district under the exclusive jurisdiction of the Congress and the District is therefore not a part of any state. The states of Maryland and Virginia each donated land to form the federal district, which included the pre-existing settlements of Georgetown and Alexandria. Named in honor of President George Washington, the City of Washington was founded in 1791 to serve as the new national capital. In 1846, Congress returned the land originally ceded by Virginia; in 1871, it created a single municipal government for the remaining portion of the District.

Washington had an estimated population of 693,972 as of July 2017, making it the 20th largest American city by population. Commuters from the surrounding Maryland and Virginia suburbs raise the city's daytime population to more than one million during the workweek. The Washington metropolitan area, of which the District is the principal city, has a population of over 6 million, the sixth-largest metropolitan statistical area in the country.

All three branches of the U.S. federal government are centered in the District: U.S. Congress (legislative), President (executive), and the U.S. Supreme Court (judicial). Washington is home to many national monuments and museums, which are primarily situated on or around the National Mall. The city hosts 177 foreign embassies as well as the headquarters of many international organizations, trade unions, non-profit, lobbying groups, and professional associations, including the Organization of American States, AARP, the National Geographic

```
Society, the Human Rights Campaign, the International Finance
Corporation, and the American Red Cross.
A locally elected mayor and a 13-member council have governed
the District since 1973. However, Congress maintains supreme
authority over the city and may overturn local laws. D.C.
residents elect a non-voting, at-large congressional delegate
to the House of Representatives, but the District has no
representation in the Senate. The District receives three
electoral votes in presidential elections as permitted by the
Twenty-third Amendment to the United States Constitution,
ratified in 1961."""
### Parse the text with spaCy
### Our 'document' variable now contains a parsed version of
text.
document = nlp(text)

### Extracting semi-structured statements
statements =
textacy.extract.semistructured_statements(document,
"Washington")

print("**** Information from Washington's Wikipedia page
****")
count = 1
for statement in statements:
    subject, verb, fact = statement
    print(str(count) + " - Statement: ", statement)
```

```
    print(str(count) + " - Fact: ", fact)
    count += 1
```

```
**** Information from Washington's Wikipedia page ****
1 - Statement:  (Washington, is, the capital of the United States of
America.[4)
1 - Fact:  the capital of the United States of America.[4
2 - Statement:  (Washington, is, the principal city of the
Washington metropolitan area, which has a population of 6,131,977.
[6)
2 - Fact:  the principal city of the Washington metropolitan area,
which has a population of 6,131,977.[6
3 - Statement:  (Washington, is, home to many national monuments and
museums, which are primarily situated on or around the National
Mall)
3 - Fact:  home to many national monuments and museums, which are
primarily situated on or around the National Mall
```

我们的NLP模型从这篇文章中发现了关于华盛顿特区的三个有用的事实：

(1)华盛顿是美国的首都

(2)华盛顿的人口，以及它是大都会的事实

(3)许多国家纪念碑和博物馆

最好的部分是，这些都是这一段文字中最重要的信息！

深入研究NLP

到这里就结束了我们对NLP的简单介绍。我们学了很多，但这只是一个小小的尝试…

NLP有许多更好的应用，例如语言翻译，聊天机器人，以及对文本文档的更具体和更复杂的分析。今天的大部分工作都是利用深度学习，特别是递归神经网络(RNNs)和长期短期记忆(LSTMs)网络来完成的。

如果你想自己玩更多的NLP，看看spaCy文档[2] 和textacy文档[3] 是一个很好的起点！你将看到许多处理解析文本的方法的示例，并从中提取非常有用的信息。所有的

东西都是快速和简单的，你可以从中得到一些非常大的价值。是时候用深入的学习来做更大更好的事情了！

参考链接：

[1] https://spacy.io/usage/linguistic-features#entity-types

[2]https://spacy.io/api/doc

[3]http://textacy.readthedocs.io/en/latest/

原文链接：

https://towardsdatascience.com/an-easy-introduction-to-natural-language-processing-b1e2801291c1

想要了解更多资讯，请扫描下方二维码，关注机器学习研究会

转自：　专知