# 第5章

# NLP应用

在这一章中，我们要来具体讨论一下NLP应用。也就是说，我们接下来会将用到在之前章节中所学到的所有概念，看看这些概念究竟能开发出何种应用程序。因此，这会是一个完全需要动手实践的章节。在前面的章节中，我们已经学习了所有NLP应用都需要执行的大部分预处理步骤。我们了解了如何使用标识器、POS标签、NER以及如何进行文本解析。本章要提供的是一种思路，让您了解应该如何运用之前所学到的知识开发出一些复杂的NLP应用。

如今，我们的现实世界中已经存在着非常多的NLP应用程序，譬如Google Search、Siri、机器翻译，Google News、Jeopardy[[1]](#footnote-1)和拼写检查等都是一些大家最为耳熟能详的例子。这其中的一些技术层次是研究人员多年努力的成果，他们将这些技术应用到了当前的水平。 NLP太复杂了，正如我们在之前章节中所看到的那样，像POS和NER这样的预处理步骤大部分也还都是研究性的问题。但通过使用NLTK库，我们已经在恰当的精确度范围内解决了其中的许多问题。我们在这本书中不会涉及到机器翻译和语音识别这样较为复杂的应用。但您现在应该已经具备了足够多的背景知识，也是时候去了解该领域的一些基本应用了。即使是作为一个NLP爱好者，我们也应该对这些NLP应用有一个基本的了解。我们也建议读者可以去互联网上找一些NLP应用来看看，并试着去了解它们。

总而言之，在本章：

* 我们将为读者介绍几个常见的NLP应用。
* 我们将会利用到目前为止所学习的知识开发一个NLP应用（新闻摘要器）。
* 我们还会介绍不同NLP应用的侧重点，以及它们各自的基本细节。

## 构建第一个NLP应用

让我们先来看一种非常复杂的NLP应用：**信息摘要（summarization）**。该应用的概念非常简单。即对于我们所提供的文章/短文/故事，您通常会需要针对其内容自动生成一些摘要。事实上，信息摘要这个应用需要我们具备一些深层次的NLP知识，因为这里需要理解的不只是句子的结构，而是整个文本的结构，除此之外，我们还得要了解该文本的体裁和主题内容。

鉴于这一切看上去都过于复杂，所以我们还是先来尝试一种很直观的方法。我们就假设这里所要做的信息摘要只不过就是根据相关句子对于我们的重要性和意义进行一次排名。我们将在理解的基础上创建一系列规则，然后用我们到目前为止所学到的处理工具来对新闻文章进行一些可接受的信息汇总处理。

在接下来的这个例子中，我们会将从*纽约时报*上搜刮来的一篇文章保存在nyt.txt这个文本文件中。在这里，我们要对这篇新闻稿进行信息摘要。下面就让我们来创建一个个人版的Google News吧。

一开始，我们首先需要记住一件事：即在通常情况下，拥有较多实体和名词的句子的重要性往往会相对比较高。现在，我们的任务是要用某种可被标准化的统一逻辑来计算**重要性评分（importance score）**。即如果我们想获取前n个句子的信息情况，就要去选择一个与其重要性评分的阈值。

现在我们来看看新闻稿的内容。在这里，您也可以选择将我这篇新闻稿以纯新闻内容的形式转储到一个文本文件中。这段内容具体如下：

>>>import sys

>>>f=open('nyt.txt','r')

>>>news\_content=f.read()

""" President Obama on Monday will ban the federal provision of some types of military-style equipment to local police departments and sharply restrict the availability of others, administration officials said.

The ban is part of Mr. Obama's push to ease tensions between law enforcement and minority communities in reaction to the crises in Baltimore; Ferguson, Mo.; and other cities. - - blic." It contains dozens of recommendations for agencies throughout the country."""

一旦我们要对这段新闻内容的解析，我们就会整个新闻稿分解成一个句子列表。这样，我们就回到了之前讨论过的句子标识器上了，后者会将整个新闻片段分解成若干个句子。在这里，我们会提供一些句型编号，便于我们识别这些句子并对其进行排名。一旦我们得到了这些句子，我们会让其在单词标识器中过一遍，最后再来过NER标注器和POS标注器。

>>>import nltk

>>>results=[] >>>for sent\_no,sentence in enumerate(nltk.sent\_tokenize(news\_content)):

>>> no\_of\_tokens=len(nltk.word\_tokenize(sentence))

>>> #print no\_of\_toekns

>>> # Let's do POS tagging

>>> tagged=nltk.pos\_tag(nltk.word\_tokenize(sentence))

>>> # Count the no of Nouns in the sentence

>>> no\_of\_nouns=len([word for word,pos in tagged if pos in ["NN","NNP"] ])

>>> #Use NER to tag the named entities.

>>> ners=nltk.ne\_chunk(nltk.pos\_tag(nltk.word\_tokenize(sentence)), binary=False)

>>> no\_of\_ners= len([chunk for chunk in ners if hasattr(chunk, 'node')])

>>> score=(no\_of\_ners+no\_of\_nouns)/float(no\_of\_toekns)

>>>

>>> results.append((sent\_no,no\_of\_tokens,no\_of\_ners,\ no\_of\_nouns,score,sentence))

在上面的代码中，我们对一个句子列表进行了迭代，并根据公式计算这些句子的评分。，该公式也只是以被标识实体为分子，以普通标识词为分母的分子式。我们会将所有的这些结果创建成一个元组。

现在，结果就是一个包含了所有评分的元组，例如其中的名词数量、实体数量等。下面我们要对评分来一个降序排序，代码如下：

>>> for sent in sorted(results,key=lambda x: x[4],reverse=True):

>>> print sent[5]

这样一来，我们就等于完成了对这些句子的排名。您会为这篇新闻稿能得到这样的结果而感到惊讶。

事实上，一旦我们手里有了no\_of\_nouns和no\_of\_ners的评分列表，就可以围绕着它们去建立一些更复杂的规则。 例如，一篇典型的新闻稿通常都会在文章的开头来一个主题说明，并在文章的最后一句也会对整个故事做一个总结。

我们可以修改之前那段代码，将上述逻辑整合进去吗？

当然，这种信息摘要应用还有另一种理论逻辑，就是重要的句子通常包含着重要的词汇，而跨语料库的差异词（discriminatory word）绝大多数都是重要的词汇。因此，只要句子中包含具有很大差异性的词汇，它就是重要的。这样，我们得到了 一个非常简单的测量方法，就是计算每个词各自的**TF-IDF（term frequency–inverse document frequency）**[[2]](#footnote-2)分值，然后根据词汇的重要性找出一种标准化的平均评分。这个评分就可以用来充当我们在信息摘要中选取句子的标准。

为了解释清楚概念，我们这里不会拿整篇文章来举例，这里将只采用文章的前三个句子。下面，我们就来看看如何用寥寥几行代码实现这个复杂的东西：

|  |
| --- |
| 这段代码需要您安装一下scikit这个库。如果您已经安装了anaconda或canopy，那么其实就已经安装了这个库，否则请按照下面链接中的指示安装scikit。  <http://scikit-learn.org/stable/install.html> |

>>> import nltk

>>> from sklearn.feature\_extraction.text import TfidfVectorizer

>>> results=[]

>>> news\_content="Mr. Obama planned to promote the effort on Monday during a visit to Camden, N.J. The ban is part of Mr. Obama's push to ease tensions between law enforcement and minority \communities in reaction to the crises in Baltimore; Ferguson, Mo. We are, without a doubt, sitting at a defining moment in American policing, Ronald L. Davis, the director of the Office of Community Oriented Policing Services at the Department of Justice, told reporters in a conference call organized by the White House"

>>> sentences=nltk.sent\_tokenize(news\_content)

>>> vectorizer = TfidfVectorizer(norm='l2', min\_df=0, use\_idf=True, smooth\_ idf=False, sublinear\_tf=True)

>>> sklearn\_binary=vectorizer.fit\_transform(sentences)

>>> print countvectorizer.get\_feature\_names()

>>> print sklearn\_binary.toarray()

>>> for i in sklearn\_binary.toarray():

>>> results.append(i.sum()/float(len(i.nonzero()[0]))

在上述代码中，我用到了一些未知的方法，譬如说TfidfVectorizer()，这是一个评分方法，它会为给定句子列表中每个句子计算出一个TF-IDF评分的向量。现在先别操心这个，我们以后会更详细地来讨论它。本章先暂且将其视为一个黑盒子函数，即对于一个给定的句子/文档列表，它将会给出每个句子所对应的评分，并且还会提供能构建出一个term-doc的矩阵，以作为我们的输出。

现在，我们从所有的句子中得到了一个容纳所有单词的字典以及一个评分列表的列表，后者的每一个元素都代表了一个单词所被赋予的TF-IDF评分。如果我们得到的是正确的结果，应该就可以看到一些停用词的评分值是接近0的，而一些差异词（譬如ban和obama）则通常会得到一个很高的评分。一旦我们在代码中获得了这些数据，就可以通过那些TF-IDF值非零的单词来得出TF-IDF的平均分值。最后我们就会得到一个与第一个方法类似的评分结果。

您一定会为这个简单的算法能得到这样的结果而感到惊讶。现在，我想我们应该已经做好了所有的准备，可以去编写属于自己的新闻摘要器了。该摘要器会利用上述两种算法对任意两篇给定的新闻搞进行信息摘要处理，并获取不错的摘要结果。虽然这种方法可以实现一个相对还不错的信息摘要程序，但它与信息摘要方面的当前研究水平相比，实际上还差得很远。我们建议读者们多去找一些信息摘要方面的文献来阅读。同时，我们也希望读者能试着混合使用这两种信息摘要的方法。

## 其它NLP应用

我们另外还有一些别的NLP应用，其中包括了文本分类、机器翻译、语音识别、信息检索、信息提取、主题划分和话语分析。这其中有一些问题其实到目前位置都还是一个非常难以实现的NLP任务，相关的领域仍在进行着大量的研究。我们将会在下一章中深入地讨论其中的一些话题，但既然是在学习NLP，我们就应该先要对这些应用的基本情况有一个了解。

### 机器翻译

对机器翻译（machine translation）最简单直接的理解方法就看我们自己是如何将某种语言翻译成另一种语言的。我们会在头脑中对相关句子的结构进行解析，以便试着理解该句子。 一旦理解了句子的含义，我们就会试着将原始语言中的单词替换成目标语言的对应词汇。并且在替换过程中，遵守目标语言的语法规则，以便最终实现正确的翻译。

（图：图中翻译

Interlingua：中间语言

Semantic composition：语义组合

Semantic decomposition：语义分解

Semantic structure：语义结构

Semantic transfer：语义转换

Semantic analysis：语义分析

Semantic generation：语义生成

Syntactic structure：句法结构

Syntactic analysis：句法分析

Syntactic transfer：句法转换

Syntactic generation：句法生成

Word structure：单词结构

Direct：直接替换

Morphological analysis：词法分析

Morphological generation：词法生成

Source text：原始文本

Target text：目标文本

）

大致上而言，翻译就是上图中那样一个金字塔状的过程。如果以原始语言的文本为出发点，我们就必须先要目标句子进行标识化处理，后者会被解析成树状结构（简而言之就是它的语法结构），以确保这些句子的正确表达。紧接着就是语义结构，这代表的这些句子所表达的含义，然后，我们来到了中间语言这个层面，该中间语言是一种独立于所有语言之外的抽象状态。到目前为止，人们已经开发出了许多种翻译方法，这些方法的走向越接近上述金字塔的塔顶，就越需要用到NLP技术。因此，根据各种不同的翻译层次，我们可以使用各种不同的方法来应对。下面我们就来看看其中的两种方法：

* **直接翻译：**这更像是一种基于字典的机器翻译，当我们拥有大型语料库以及海量的目标语言词汇时，依赖于相关语言的大型语料库来实现某种类型的翻译应用是有可能的。而且因它简单而流行。
* **语法翻译：**在这种翻译方法中，我们会试着去构建一个针对原始语言的解析器。到目前为止，人们已经累积了各种各样的关于解析问题的方法。其中有些深层解析器实际上已经拥有了处理一部分语义的能力。一旦我们搞定了解析器，目标词汇的替换问题就迎刃而解了，目标解析器会自行产生出最终的目标语言的句子。

### 统计型机器翻译

统计型机器翻译（**Statistical machine translation** ，简称SMT）是一种最新型的机器翻译方法。在这种方法中，人们提出了各种运用统计学方法，几乎涵盖了机器翻译方面的所有面向。 这一类算法背后的思路是依靠我们所拥有的海量语料库、并行化文本以及可用目标语言产生语言翻译的语言模型。Google Translate就是一个很好的SMT应用实例，它会从不同语言的语料库中学习到相关信息，并围绕这些信息来创建SMT应用。

### 信息检索

**Information retrieval** (**IR**) is also one of the most popular and widely used applications. The best exmple of IR is Google Search, where—given an input query from the user—the information retrieval algorithm will try to retrieve the information which is relevant to the user's query.

信息检索（IR）也是最受欢迎和广泛使用的应用之一。 IR的最好例子是Google搜索，在给定来自用户的输入查询的情况下，信息检索算法将尝试检索与用户的查询相关的信息。

In simple words, IR is the process of obtaining the most relevant information that is needed by the user. There are a variety of ways in which the information needs can be addressed to the system, but the system eventually retrieves the most relevant infromation.

简单地说，IR是获得用户所需的最相关的信息的过程。 有多种方式可以将信息需求寻址到系统，但是系统最终检索到最相关的信息。

The way a typical IR system works is that it generates an indexing mechanism, also known as **inverted index**. This is very similar to the indexing schemes used in books, where you will have an index of the words present throughout the book on the last pages of the book. Similarly, an IR system will create an inverted index poslist. A typical posting list will look like this:

典型的IR系统工作的方式是它生成索引机制，也称为反向索引。 这与书中使用的索引方案非常相似，您将在书的最后一页上找到整本书中出现的单词的索引。 类似地，IR系统将创建反向索引poslist。 典型的过帐列表将如下所示：

< Term , DocFreq, [DocId1,DocId2] >

{"the",2 --->[1,2] }

{"US",1 --->[2] }

{"president",2 --->[1,2] }

So if any word occurs in both document 1 and document 2, the posting list will be a list of documents pointing to terms. Once you have this kind of data structure, there are different retrieval models that can been introduced. There are different retrieval models that work on different types of data. A few are listed in the following sections.

因此，如果在文档1和文档2中出现任何单词，则发布列表将是指向词语的文档的列表。 一旦你有这种数据结构，就有可能引入不同的检索模型。 有不同的检索模型，工作在不同类型的数据。 以下部分列出了几个。

#### Boolean retrieval

In the Boolean model, we just need to run a Boolean operation on the poslist. For example, if we are looking for a search query like "US president", the system should look for an intersection of the postlist of "US" and "president".

在布尔模型中，我们只需要在poslist上运行一个布尔运算。 例如，如果我们正在寻找类似“美国总统”的搜索查询，系统应该查找“美国”和“总统”的后续列表的交集。

{US}{president}=> [2]

Here, the second document turns out to be the relevant document.

这里，第二文档证明是相关文档。

#### Vector space model

The concept of **vector space model** (**VSM**) derives from geometry. The way to visualize the documents in the high dimension space of vocabulary is to represent it as a vector. So each and every document is represented as a vector in that space. We can represent the vector in various ways, but one of the most useful and efficient ways is using TF-IDF.

向量空间模型（VSM）的概念来自几何。 在词汇的高维空间中可视化文档的方式是将其表示为向量。 因此，每个文档都表示为该空间中的一个向量。 我们可以用各种方式表示向量，但最有用和最有效的方法之一是使用TF-IDF。

Given a term and a corpus, we can calculate the **term frequency** (**TF**) and **inverse document frequency** (**IDF**) using the following formula:

给定一个项和语料库，我们可以使用以下公式计算项频率（TF）和逆文档频率（IDF）：

（公式）

The TF is nothing but the frequency in the document. While the IDF is the inverse of document frequency, which is the count of documents in the corpus where the term occurs:

TF不过是文档中的频率。 虽然IDF是文档频率的倒数，其是语料库中出现术语的文档的计数：

（公式）

There are various normalization variants of these, but we can incorporate both of these to create a more robust scoring mechanism to get the scoring of each term in the document. To get to a TF-IDF score, we need to multiply these two scores as follows: tfidf (*t d*, ,*D*)= tf (*t*,*d*)×idf (*t*,*D*)

有这些的各种标准化变体，但我们可以合并这两个，以创建一个更可靠的评分机制，得到文档中每个术语的评分。 为了得到TF-IDF分数，我们需要将这两个分数相乘如下：tfidf（t d，D）= tf（t，d）×idf（t，D）

In TF-IDF, we are scoring a term for how much it is present in the current document and how much it is spread across the corpus. This gives us an idea of the terms that are not common across corpora and where ever they are present have a high frequency. It becomes discriminatory to retrieve these documents. We have also used TF-IDF in the previous section, where we describe our summarizer.The same scoring can be used to represent the document as a vector. Once we have all the documents represented in a vectorized form, the vector space model can be formulated.

在TF-IDF中，我们对当前文档中存在多少以及它在语料库中传播多少进行评分。 这给了我们一个想法，不是常见的语料库，而且他们存在的术语有一个高频率。 检索这些文档变得有歧视性。 我们还在前面的部分中使用了TF-IDF，我们在其中描述我们的摘要器。相同的评分可以用于将文档表示为向量。 一旦我们得到以向量化形式表示的所有文档，就可以制定向量空间模型。

In VSM, the search query of the user is also considered as a document and represented as a vector. Intuitively, a dot product between these two vectors can be used to get the cosine similarity between the document and the user query.

在VSM中，用户的搜索查询也被认为是文档并被表示为向量。 直观地，这两个向量之间的点积可以用于获得文档和用户查询之间的余弦相似性。

（图：图中翻译）

In the preceding diagram, we see that these same documents can be represented using each term as an axis and the query Obama will have as much relevance to *D1* as compared to *D2*. The scoring of the query for relevant documents can be formulated as follows:

在上图中，我们看到这些相同的文档可以使用每个术语作为轴来表示，并且查询Obama将与D1相比具有与D2相同的相关性。 对相关文件的查询得分可以表示如下：

（公式）

#### The probabilistic model

The probabilistic model tries to estimate the probability of the user's need for the document. This model assumes that the probability of the relevance depends on the user query and document representation.The main idea is that a document that is in the relevant set will not be present in the non-relevant set. We denote dj as the document and q as user query; R represents the relevant set of documents, while P represents the non-relevant set. The scoring can be done like this:

概率模型尝试估计用户对文档的需要的概率。 该模型假定相关性的概率取决于用户查询和文档表示。主要思想是在相关集合中的文档将不存在于非相关集合中。 我们将dj表示为文档，q表示为用户查询; R表示文档的相关集合，而P表示非相关集合。 评分可以这样做：

（公式）

|  |
| --- |
| For more topics on IR, I would recommend that you read from the following link:  http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html |

### 语音识别

Speech recognition is a very old NLP problem. People have been trying to address this since the era of World War I, and it still is one of the hottest topics in the area of computing. The idea here is really intuitive. Given the speech uttered by a human can we convert it to text? The problem with speech is that we produce a sequence of sounds, called **phonemes**, that are hard to process, so speech segmentation itself is a big problem. Once the speech is processable, the next step is to go through some of the constraints (models) that are built using training data available. This involves heavy machine learning. If you see the figure representing the modeling as one box of applying constraints, it's actually one of the most complex components of the entire system. While acoustic modeling involves building modes based on phonemes, lexical models will try to address the modeling on smaller segments of sentences, associating a meaning to each segment. Separately language models are built on unigrams and bigrams of words.

语音识别是一个非常老的NLP问题。自从第一次世界大战时代以来，人们一直在努力解决这个问题，它仍然是计算领域最热门的话题之一。这里的想法真的很直观。给定人类发出的语音，我们可以将其转换为文本吗？语音的问题是我们产生一系列被称为音素的难以处理的声音，因此语音分割本身是一个大问题。一旦语音是可处理的，下一步是通过使用可用的训练数据构建的一些约束（模型）。这涉及重型机器学习。如果您将该图表示为一个应用约束的框，它实际上是整个系统中最复杂的组件之一。虽然声学建模涉及基于音素建立模式，但词汇模型将尝试解决对较小句子段的建模，将意义与每个段相关联。单独的语言模型建立在单字和双字的基础上。

Once we build these models, an utterence of the sentences is passed through the process. Once processed for initial preprocessing, the sentence is passed through these acoustic, lexical, and language models for generating the token as output.

一旦我们构建这些模型，句子的语句通过该过程。一旦被处理用于初始预处理，句子被传递通过这些声学，词汇和语言模型用于生成令牌作为输出。

（图：图中翻译）

### 文本分类

Text classification is a very interesting and common application of NLP. In your daily work, you interact with many text classifiers. We use a spam filter, a priority inbox, news aggregators, and so on. All of these are in fact applications built using text classification.

文本分类是NLP非常有趣和常见的应用。 在你的日常工作中，你与许多文本分类器交互。 我们使用垃圾邮件过滤器，优先收件箱，新闻聚合器等。 所有这些都是使用文本分类构建的应用程序。

Text classification is a well-defined and somewhat solved problem, and it has been applied across many domains. Typically, any text classification is the process of classifying text documents using words and the combination of words. While it's a typical machine learning problem, many of the preprocessing steps used in text classification are from NLP.

文本分类是一个定义明确且有些解决的问题，它已经应用于许多领域。 通常，任何文本分类是使用词和词的组合来分类文本文档的过程。 虽然这是一个典型的机器学习问题，文本分类中使用的许多预处理步骤都来自NLP。

An abstract diagram of text classification is shown here:

文本分类的抽象图如下所示：

（图：图中翻译）

Here we have a bunch of documents for a set of classes. For simplicity, we will use just binary 1/0 as the class. Now let's assume it's a spam detection problem where 1 represents spam and 0 represents normal text which is not to be considered as spam.

这里我们有一堆文档的一组类。为了简单起见，我们只使用二进制的1/0作为类。现在让我们假设这是一个垃圾邮件检测问题，其中1表示垃圾邮件，0表示不被视为垃圾邮件的正常文本。

The process involves some of the preprocessing steps we learned in previous chapters. While some of these are essential, it depends on the kind of text classification problem we are trying to solve. So in few cases, it's more a case of feature engineering while we drop some of the preprocessing steps. The final goal of feature engineering is to generate a **Term doc matrix** (**TDM**), which holds the vocabulary of the entire corpus: columns and rows are the documents, while the matrix represents a scoring mechanism to show the **Bag of word** (**BOW**) representation. The weighting scheme can be varied to TF, TF-IDF, Bernoulli, and other variations of term frequency.

该过程涉及我们在前面章节中学到的一些预处理步骤。虽然其中一些是必要的，它取决于我们试图解决的文本分类问题的种类。所以在少数情况下，这是更多的特征工程的情况，而我们放弃一些预处理步骤。特征工程的最终目标是生成一个术语文档矩阵（TDM），其保存整个语料库的词汇表：列和行是文档，而矩阵表示用于显示词袋（BOW）表示的评分机制。加权方案可以改变为TF，TF-IDF，伯努利和术语频率的其他变化。

There are also ways to induce features such as the POS of a given feature, contextual POS, and others, to make our feature space more NLP intense. Once the

还有一些方法来诱导特征，例如给定特征的POS，上下文POS以及其他特征，以使我们的特征空间更加NLP激烈。一旦

TDM is generated, the text classification problem becomes a typical supervised/ unsupervised classification problem, where given a set of samples, we need to predict what sample belongs to what class. The next chapter is dedicated entirely to this topic. This is definitely a splendid application of NLP/ML and is used quite often for commercial purposes.

TDM生成，文本分类问题变成典型的监督/非监督分类问题，其中给定一组样本，我们需要预测什么样本属于什么类。下一章专门讨论这个话题。这绝对是NLP / ML的一个辉煌的应用，并且经常用于商业目的。

Some of the most common use cases in day-to-day scenarios are sentiment analysis, spam classification, e-mail categorization, news categorization, patent classification, and so on. We will talk about text classification in more detail in the next chapter.

在日常情况下，一些最常见的用例是情绪分析，垃圾邮件分类，电子邮件分类，新闻分类，专利分类等。我们将在下一章更详细地讨论文本分类。

### 信息提取

**Information extraction** (**IE**) is a process of extracting meaningful information from unstructured text. IE is yet another widely popular and highly important application. In general, an information extraction engine harnesses huge numbers of unstructured documents and generates some sort of structured/semi-structured **knowledge base** (**KB**) that can be deployed to build an application around it. A simple example is that of generating a very good ontology using a huge set of unstructured text documents. A similar project in this line is DBpedia, where all the Wikipedia articles are used to generate the ontology of artifacts that are interrelated or have some other relationship.

信息提取（IE）是从非结构化文本中提取有意义的信息的过程。 IE是另一个广泛流行和非常重要的应用程序。一般来说，信息提取引擎利用大量非结构化文档，并生成某种结构化/半结构化知识库（KB），可以部署该知识库以在其周围构建应用程序。一个简单的例子是使用一组巨大的非结构化文本文档生成一个非常好的本体。在这一行中的一个类似的项目是DBpedia，其中所有的维基百科文章用于生成相互关联或有一些其他关系的工件的本体。

There are mainly two ways of extracting information:

主要有两种提取信息的方法：

* **Rule-based extraction**: This method is where one uses a template filling mechanism. The idea is to look for some kind predefined use cases for expected outcomes and try to mine the unstructured text for that specific template. For example, building a knowledge base of football will involve getting information on all the players and their profiles, the statistics, some personal information, and so on. All that can be well defined and extracted using either pattern-based rules or POS tags, NERs and relation extraction.
* **Machine learning based**: The other approach involves deeper NLP-based methods such as building a parser specific to the need of our knowledge base. Some of the KBs will require mining the entities that can't be extracted using a pre-trained NER, so we have to build a custom NER. We might want to develop a relation extraction algorithm specific to the KB we are trying to build. This is a more NLP-intensive approach, where we are developing a NLP-based parser or tagger to use for heavy machine learning.
* 基于规则的提取：这种方法是使用模板填充机制。这个想法是为预期的结果寻找一些预定义的用例，并尝试挖掘该特定模板的非结构化文本。例如，构建足球知识库将涉及获取关于所有玩家及其简档，统计数据，一些个人信息等的信息。所有这些都可以很好地定义和提取使用基于模式的规则或POS标签，NER和关系提取。
* 基于机器学习：另一种方法涉及更深层的基于NLP的方法，例如构建一个专门针对知识库需求的解析器​​。一些知识库将需要挖掘不能使用预训练的NER提取的实体，所以我们必须建立一个自定义的NER。我们可能想要开发一个特定于我们正在尝试构建的KB的关系提取算法。这是一种更加NLP密集型的方法，我们正在开发一个基于NLP的解析器或标记器，用于重型机器学习。

### 问答系统

**Question answering** (**QA**) systems are intelligent systems that can address any question based on their knowledge base. One of the major examples of this is IBM Watson, which took part in the TV show *Jeopardy* and won over human opponents. A QA system can be broken down to building components from speech recognition for querying the knowledge base while the knowledge base is generated using information retrieval and extraction.

问答（QA）系统是可以基于其知识库解决任何问题的智能系统。 其中一个主要的例子是IBM Watson，他参加了电视节目Jeopardy，赢得了人类对手。 QA系统可以被分解为从用于查询知识库的语音识别构建组件，而使用信息检索和提取来生成知识库。

Once you have a question for the system, one big problem is to classify/categorize the question in different ways. The other aspect is to search the knowledge base effectively and retrieve the most precise document. Even after that, we have to generate the answer in a natural way using some of the other applications, such as summarization and parsing.

一旦你对系统有一个问题，一个大问题是以不同的方式对问题进行分类/分类。 另一方面是有效地搜索知识库并检索最精确的文档。 即使之后，我们必须使用一些其他应用程序以自然的方式生成答案，例如摘要和解析。

### 对话系统

Dialog systems are considered the dream application, where given a speech in source language, the system will perform speech recognition and transcribe it to text. This text will then go to a machine translation system that can translate the speech into the target language and then a text-to-speech system will convert it into speech in the target language. This is one of the most desirable applications of NLP, where we can talk to a computer in any language and the computer will reply in the same language. This kind of application can actually destroy the language barrier that exists in the world.

Apple Siri and Google Voice are examples of some of the commercial applications in the line of dialog systems intelligent enough to understand our information needs, try to address them in a set of actions or information, and respond in a human-like manner.

Apple Siri和Google Voice是一些对话系统中的一些商业应用程序的例子，足够了解我们的信息需求，尝试在一系列动作或信息中解决它们，并以类似人的方式做出响应。

### 词义消歧

**Word sense disambiguation** (**WSD**) is also one of the difficult challenges not solved even after years of research and one of the major causes of application problems, such as question answering, summarization, search, and so on. A simple way to understand the concept is that many words have different meanings when used in different contexts. For example, "cold" in the following example:

词义消歧（WSD）也是即使在多年的研究和应用问题的主要原因之一未解决的困难挑战之一，例如问题回答，摘要，搜索等。 一个理解概念的简单方法是，当在不同的上下文中使用时，许多单词具有不同的含义。 例如，“cold”在以下示例中：

* The ice-cream is really cold
* That was cold blooded!
* 冰淇淋真的很冷
* 那是冷血！

Here the word "cold "has two different senses, and it's really hard for computers to understand this concept. Some of the other NLP processing options, such as POS tagging and NER, are used to resolve some of these problems.

这里词“冷”有两种不同的感觉，计算机真的很难理解这个概念。 一些其他NLP处理选项，如POS标记和NER，用于解决这些问题中的一些。

### 主题模型

Topic modeling, in the context of a large amount of unstructured text content, is really an amazing application, where the primary task is to identify the emerging topics in the corpus and then categorize the documents in the corpus as per these topics. We will discuss this briefly in the next chapter.

在大量非结构化文本内容的上下文中，主题建模真的是一个了不起的应用程序，其中主要任务是识别语料库中新出现的主题，然后根据这些主题将文档分类在语料库中。 我们将在下一章中简要讨论这个问题。

Topic modeling uses the same NLP preprocessing, for example, sentence split, tokenization, stemming, and so on. The beauty of the algorithms is that we have an unsupervised way of categorizing the document; also, topics are generated without explicitly mentioning anything prior to the process. I encourage you to look at topic modeling in more detail. Try reading about **latent dirichlet allocation** (**LDA**) and **latent semantics indexing** (**LSI**) for more detail.

主题建模使用相同的NLP预处理，例如句子拆分，标记化，词干化等。 算法的优点是我们有一种无人监管的文档分类方法; 同样，生成主题而没有在处理之前明确提及任何东西。 我鼓励你更详细地看主题建模。 尝试阅读有关潜在狄利克雷分配（LDA）和潜在语义索引（LSI）的更多详细信息。

### 语言检测

Given a snippet of text, the detection of language is also a problem. The application of language detection is very important for some of the other NLP applications, such as search, machine translation, speech, and so on. The main concept is learning from the text as features what the language is. A variety of machine learning and NLP techniques are used for feature engineering in the process.

给定一个文本片段，检测语言也是一个问题。 语言检测的应用对于其他一些NLP应用程序非常重要，例如搜索，机器翻译，语音等。 主要的概念是从文本学习作为特征什么是语言。 各种机器学习和NLP技术用于过程中的特征工程。

### 光符识别

**Optical character recognition** (**OCR**) is an application of NLP and computer vision, where given a handwritten document/ non-digital document, the system can recognize the text and extract it into digital format. This has also been widely researched in the area of machine learning for many years. Some of the big OCR projects are Google Books, where they use OCR to convert non-digital books into a centralized library.

光学字符识别（OCR）是NLP和计算机视觉的应用，其中给定手写文档/非数字文档，系统可以识别文本并将其提取为数字格式。 这也已经在机器学习领域中被广泛研究多年。 一些大型OCR项目是Google图书，他们使用OCR将非数字图书转换为集中式图书馆。

## 本章小结

In conclusion, there are many NLP applications around us that we interact with in our day-to-day routines. NLP is difficult and complex, and some of these problems are still unsolved or do not yet have perfect solutions. So anybody who is looking for problems in NLP, try exploring the literature around that. It's a great time to be an NLP researcher. In the era of Big Data, NLP applications are very popular. Many research labs and organizations are currently working on NLP applications such as speech recognition, search, and text classification.

总之，在我们周围有很多NLP应用程序，我们在我们的日常例程中进行交互。 NLP是困难和复杂的，并且这些问题中的一些仍然没有解决或还没有完美的解决方案。所以任何人谁在寻找NLP中的问题，尝试探索这方面的文献。现在是成为一名NLP研究员的好时机。在大数据时代，NLP应用程序非常受欢迎。许多研究实验室和组织目前正在开发NLP应用程序，如语音识别，搜索和文本分类。

I believe we have learned a lot up until this chapter. For the next couple of chapters, we will delve deeply into some of the applications described here. We have reached a point where we know enough NLP related preprocessing tools and also have a basic understanding about some of the most popular NLP applications. I hope you leverage some of this learning to build a version of an NLP application.

In the next chapter, we will start with some of the important NLP applications, such as text classification, text clustering, and topic modeling. We will move slightly away from the pure NLTK applications on to how NLTK can be used in conjunction with other libraries.

在下一章中，我们将从一些重要的NLP应用程序开始，如文本分类，文本聚类和主题建模。我们将稍微离开纯NLTK应用程序，了解NLTK如何与其他库结合使用。

1. 译者注：这是一款文字类问答游戏，这些问答题非常考验玩家的英文水平、以及各个领域的知识。玩家要有能力解析题目中的隐晦含义，反讽或者谜题。这也是目前计算机最欠缺的能力。 [↑](#footnote-ref-1)
2. 译者注：TF-IDF是一种统计方法，主要用来评估某一单词在一个文件集或某个语料库中某一份文件的重要程度。单词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。 [↑](#footnote-ref-2)