# 第10章 **大规模文本挖掘**

在本章，我们打算再回头来谈谈在之前章节中提到的一些程序库，但这回要谈的是如何在大数据环境中扩大规模地使用这些库。因此，我们会假设读者对于Hadoop+Hive这样的大数据框架已经有了一定的了解。在此基础之上，我们会对一些Python库进行一些相应的探讨，例如NLTK、scikit-learn和pandas这几个库都可以被应用于带有大规模非结构化数据的Hadoop集群。

我们将会讨论到一些讨论NLP和文本挖掘领域中常见的用例，在这过程中，我们也会给出一些代码片段，以便帮助您完成相关的工作。具体来说，我们要来看三个会涉及到绝大多数文本挖掘问题的主要示例。通过这些示例，我们会告诉您如何通过大规模地执行NLTK来完成本书最初几章中所介绍的那些NLP任务。此外，我们还将通过几个例子来介绍如何在大数据条件下执行文本分类任务。

当然，机器学习和NLP还有另一高度规模化应用的问题就是它们是否可并行化。我们在这里将会简单地讨论一下上一章中的一些问题，看看这些问题是否属于大数据问题。或者是否在某些条件下可以用大数据的方式来解决这些问题。

由于我们到目前为止所学习的大多数库都是用Python编写的，所以如何用Python（Hadoop）来处理大数据也是本章的主要问题之一。

在阅读完本章之后，我们希望读者：

* 能很好地了解Hadoop，Hive这些与大数据相关的技术。并在其条件下使用Python。
* 根据教程一步一步地掌握如何在大数据条件下使用NLTK、Scikit和PySpark。

## 在Hadoop上使用Python的不同方式

在Hadoop上运行一个Python进程的方式有很多种。在这里，我们将会讨论其中一些当前最为流行的方式，并通过这些方式在Hadoop上用Python来实现流式的MapReduce作业[[1]](#footnote-1)、Hive中的Python UDF、以及Python hadoop包装器。

### Python的流操作

通常，一个典型的Hadoop作业必须要被写成map+reduce函数的形式。用户需要根据给定任务来编写相应的map+reduce函数的实现。这些mapper和reducer通常是用Java来实现的。而与此同时Hadoop也为我们提供流操作的接口，用户可以基于这些接口来写一个Python封装器，并用其它任意一种语言来编写之前由Java所实现的mapper和reducer函数。接下来，我们会来看一个用Python编写的单词计数示例。而且本章稍后还会介绍如何用NLTK库再实现一次。

|  |
| --- |
| 如果您还不太了解情况，可以看看下面链接中的资料：  <http://www.michael-noll.com/tutorials/writingan-hadoop-mapreduce-program-in-python/>  了解一下Python环境下的MapReduce模式。. |

### Hive/Pig下的UDF

另一种使用Python处理大数据的方式就是在Hive / Pig中编写**UDF（User Defined Function）**。 这种方法的思路认为：我们在NLTK中所执行的大多数操作都是高度可并行化的。譬如说词性标注、标识化处理、词形还原、停用词移除以及NER这些都是可高度分布式执行的操作。因为其中的每一行内容都独立于其它行，所以我们在执行这些操作时不需要根据任何上下文。

因此，如果我们在集群上的每个节点上都部署了NLTK及其它Python库，也可以用Python来编写一些**用户定义函数（UDF）**，以借助NLTK和scikit这些库的功能。这是其引用NLTK最简单的一种方法，对于scikit的大规模引用则更是如此。我们会在本章后续内容中具体介绍这两个库的情况。

### 流封装器

我们可以将不同组织所实现的各种封装器列成一份长长的列表，以便能让人明白Python在集群上可以执行的各项任务。这其中有一些封装使用起来其实相当简单，但问题是它们都有性能较差这个问题。我在下面也列出了一些，如果您想了解它们，可以去这些项目的网站去阅读一下相关介绍：

* Hadoopy
* Pydoop
* Dumbo
* Mrjob

|  |
| --- |
| 如果想查看一份更详细的目前Hadoop上可供选择的Python库列表，读者可以参阅下面这篇文章：  <http://blog.cloudera.com/blog/2013/01/a-guide-topython-frameworks-for-hadoop/> |

## Hadoop上的NLTK

我们之前已经从一个库的角度对NLTK进行了充分的讨论，并介绍了它的一只鹅最常用的函数。目前，NLTK已经可以解决许多NLP问题，这其中有许多都是高度可并行化的方案。这也是为什么我们要试着在Hadoop上使用NLTK的原因。

在Hadoop上，运行NLTK的最佳途径就是将其安装在集群的所有节点上。这实现起来并不困难。有几种方式都可以做到这一点，例如我们可以将资源文件以流参数的形式来发送。但通常我们宁愿选择下面的第一个选项。

### 用户定义函数（UDF）

在Hadoop上运行NLTK的方法有很多，我们下面要讨论的实例是如何用NLTK来实现一个Hive UDF，以便完成并行式的标识化处理。

这个用例的操作可以分成以下步骤：

1. 我们在这里选择了一个只包含两列数据的小型数据集，然后在Hive中创建与之相对于的相同数据模式[[2]](#footnote-2)：

|  |  |
| --- | --- |
| **ID** | **内容** |
| UA0001 | "I tried calling you. The service was not up to the mark" |
| UA0002 | "Can you please update my phone no" |
| UA0003 | "Really bad experience" |
| UA0004 | "I am looking for an iPhone" |

1. 现在，我们在Hive中创建相同的数据模式，该操作的Hive脚本如下：

Hive script

CREATE TABLE $InputTableName (

ID String,

Content String

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

1. 数据模式建构完成之后，基本上，下一步要做的就是针对每一个独立列中的内容进行类似于标识化的处理。所以，我们现在就会想在$ outTable中设置另一个具有相同模式的列，并为其添加相应的标识列：

Hive script

CREATE TABLE $OutTableName (

ID String,

Content String,

Tokens String

)

1. 现在模式方面的准备已经完成了，接下来我们要用Python来编写UDF，以逐行读取上面的数据表，并对其执行tokenize()方法。这一过程与我们在第3章“词性标注”所做的事情非常类似。这是一个与第3章“词性标注”中所有示例都很类似的函数。而且，如果您现在还想获得其POS标签、词形但愿以及某HTML标记的话，只需要相应地修改一下这个UDF即可。下面我们就来看看这个UDF是如何查找相关标识的：

>>> import sys

>>> import datetime

>>> import pickle

>>> import nltk

>>> nltk.download('punkt')

>>> for line in sys.stdin:

>>> line = line.strip()

>>> print>>sys.stderr, line

>>> id, content= line.split('\t')

>>> print>>sys.stderr,tok.tokenize(content)

>>> tokens =nltk.word\_tokenize(concat\_all\_text)

>>> print '\t'.join([id,content,tokens])

1. 接下来，我们要命名一下这个UDF ，譬如nltk\_scoring.py。
2. 现在，我们得要用TRANSFORM函数来执行Hive的insert查询，以便将上面的UDF应用到给定的内容中，将新列标记化并转储其标识：

Hive script

add FILE nltk\_scoring.py;

add FILE english.pickle; #Adding file to DistributedCache

INSERT OVERWRITE TABLE $OutTableName

SELECT

TRANSFORM (id, content)

USING 'PYTHONPATH nltk\_scoring.py'

AS (id string, content string, tokens string )

FROM $InputTablename;

1. 如果您在上述过程中遇到了下面这样的错误信息，就说明NLTK机器数据组件没有被正确安装：

raiseLookupError(resource\_not\_found) LookupError:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Resource u'tokenizers/punkt/english.pickle' not found. Please use the NLTK Downloader to

obtain the resource:

>>> nltk.download() Searched in:

'/home/nltk\_data'

'/usr/share/nltk\_data'

'/usr/local/share/nltk\_data'

'/usr/lib/nltk\_data'

'/usr/local/lib/nltk\_data'

1. 如果该Hive作业能够被成功运行，我们就会得到一张名为OutTableName的表，具体如下：

|  |  |  |
| --- | --- | --- |
| **ID** | **内容** |  |
| UA0001 | "I tried calling you, The service was not up to the mark" | [" I", " tried", "calling", "you", "The",  "service" "was", "not", "up", "to", "the", "mark"] |
| UA0002 | "Can you please update my phone no" | ["Can", "you", "please" "update", " my",  "phone" "no"] |
| UA0003 | "Really bad experience" | ["Really"," bad" "experience"] |
| UA0004 | "I am looking for an iphone" | ["I", "am", "looking", "for", "an", "iPhone"] |

### Python的流操作

下面我们来试试第二个选项：Python流。我们根据现有的Hadoop流来编写自己的mapper和reducer函数，然后使用mapper.py来操作Python流，它的用法和Hive UDF非常类似。下面，我们就用同一个例子来示范一下map-reduce函数操作Python流的过程。这种方法为我们提供了一种可以使用Hive表，甚至直接使用HDFS文件的选项。在这里，我们只示范如何读取相关的内容并将其标识化，并不会涉及到任何reduce操作，但为了让学习有一定的完整性，我会在这里纳入一个虚拟的reducer，该函数只负责转储相关结果。也正因为如此，我们在这里可以完全忽略掉来自命令执行环境中的reducer。 下面是Mapper.py的代码：

Mapper.py

**>>>import sys**

**>>>import pickle**

**>>>import nltk >>>for line in sys.stdin:**

**>>> line = line.strip()**

**>>> id, content = line.split('\t')**

**>>> tokens =nltk.word\_tokenize(concat\_all\_text)**

**>>> print '\t'.join([id,content,topics])**

接下来是Reducer.py的代码：

Reducer.py

**>>>import sys**

**>>>import pickle**

**>>>import nltk >>>for line in sys.stdin:**

**>>> line = line.strip()**

**>>> id, content,tokens = line.split('\t')**

**>>> print '\t'.join([id,content,tokens])**

下面再来看看Hadoop中执行Python流操作的命令：

Hive script

hadoop jar <path>/hadoop-streaming.jar \ -D mapred.reduce.tasks=1 -file <path>/mapper.py \

-mapper <path>/mapper.py \

-file <path>/reducer.py \

-reducer <path>/reducer.py \

-input /hdfspath/infile \

-output outfile

## Hadoop 上的Scikit-learn

机器学习是大数据领域中另一个非常重要的用例。在这方面Hadoop、scikit-learn这些框架就显得更尤为重要了，因为这是我们在大数据环境下对机器学习模型进行评分的最佳选项之一。由于大型机器学习本身就是当前业界最热门的话题之一，所以在Hadoop这样的大数据环境中做这些事就显得更为重要了。当前，机器学习模型主要有两个方面的问题：在大数据环境下建模，以及针对大量数据的建模和评分。

为了加深对上述概念的理解，我们在这里也将采用之前那张表中的相同示例数据，这其中包含了一些客户的评论信息。现在假设我们要用一份重要的训练样本来构建一个文本分类模式，并利用在*第6章“文本分类”*中学到的知识在该数据上构建出朴素贝叶斯算法、SVM或逻辑回归模型。然后在进行评分的时候，我们可能需要面对的是针对海量数据的评分，譬如说客户评论这样的信息。另一方面，靠scikit-learn库本身来在大数据环境中建模是做不到的，所以我们在这里会需要用到像spark / Mahot这样的工具。另外，我们还会像在使用NLTK时一样用到某种预训练模型，并执行相同的步骤评分方法，我们将会在下一节中介绍其建模过程。特别在处理处理文本挖掘类问题时，使用预训练模型来执行评分是尤为合适的。当然，我们需要将这里的两个主要对象（vectorizer和modelclassifier）存储成某种序列化的pickle对象。

|  |
| --- |
| 这里提到的pickle其实是一个Python模块，主要用于实现序列化，即通过该模块，我们可以将对象以二进制状态存储在磁盘上，并且可以通过二次加载来使用。  <https://docs.python.org/2/library/pickle.html> |

下面，我们要用scikit在本地机器上构建离线模型，并准备好pickle对象。举例来说，如果我们像这里重现的是*第6章“文本分类”*中朴素贝叶斯算法的那个例子，就必须要将其中的vectorizer和clf转储成pickle对象：

>>> vectorizer = TfidfVectorizer(sublinear\_tf=True,

min\_df=in\_min\_df, stop\_words='english', ngram\_range=(1,2), max\_df=in\_max\_df)

>>> joblib.dump(vectorizer, "vectorizer.pkl", compress=3)

>>> clf = GaussianNB().fit(X\_train,y\_train)

>>> joblib.dump(clf, "classifier.pkl")

以下是我们创建输出表的步骤，该表中包含了所有客户评价的历史记录：

1. 在Hive中创建一个与之前示例相同的模式。下面我们就用Hive脚本来执行这一操作。这个输出表可能是很巨大的，具体到眼下这个例子，我们会假设这其中包含了公司收到过的所有客户评论：

Hive script

CREATE TABLE $InputTableName (

ID String,

Content String

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

1. 构建组成输出表的各列，例如评分可信度等：

Hive script

CREATE TABLE $OutTableName (

ID String, Content String, predict String, predict\_score double )

1. 接下来，我们要用Hive中的addFILE命令将这些pickle对象加载到其分布式缓存中：

add FILE vectorizer.pkl;

add FILE classifier.pkl;

1. 下一步就是编写Hive UDF了，我们会在这些函数中定义如何加载这些pickle对象。现在，这些对象的行为必须要先与其在本地的行为保持一致。在我们设置好这些classifier和vectorizer对象之后，接下来就可以使用我们的测试样本了。在这里，测试样本只是一个字符串，然后生成输出的是TFIDF向量。vectorizer对象现在属于预测类，可以当作概率类来使用：

Classification.py

**>>>import sys**

**>>>import pickle**

**>>>import sklearn**

**>>>from sklearn.externals import joblib**

**>>>clf = joblib.load('classifier.pkl')**

**>>>vectorizer = joblib.load('vectorizer.pkl')**

**>>>for line in sys.stdin:**

**>>> line = line.strip()**

**>>> id, content= line.split('\t')**

**>>> X\_test = vectorizer.transform([str(content)])**

**>>> prob = clf.predict\_proba(X\_test)**

**>>> pred = clf.predict(X\_test)**

**>>> prob\_score =prob[:,1]**

**>>> print '\t'.join([id, content,pred,prob\_score])**

1. 在编写完classification.py中的UDF之后，我们还必须要将这个UDF添加到分布式缓存中，然后有效地在输出表的每一行上将这个UDF以TRANSFORM函数的形式运行起来。其Hive脚本如下：

Hive script

add FILE classification.py;

INSERT OVERWRITE TABLE $OutTableName

SELECT

TRANSFORM (id, content)

USING 'python2.7 classification.py'

AS (id string, scorestringscore string )

FROM $Tablename;

1. 如果一切顺利，我们就会得到符合以下输出模式的输出表：

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **内容** | **预测** | **评分可信度** |
| UA0001 | "I tried calling you, The service was not up to the mark" | 投诉 | 0.98 |
| UA0002 | "Can you please update my phone no " | 非投诉 | 0.23 |
| UA0003 | "Really bad experience" | 投诉 | 0..97 |
| UA0004 | "I am looking for an iPhone " | 非投诉 | 0.01 |

如您所见，我们的输出表中包含了过去所有的客户评论记录，以及这些评论是否属于投诉的预测和预测的可信度。当然在该例子中我们采用的是Hive UDF，但类似的过程也可以用Pig和Python流的方式来做，其做法与用NLTK非常类似。

这个例子所演示的是如何在Hive上对机器学习模型进行评分的实践经验。 在下一个例子中，我们将要来讨论如何在大数据环境中构建机器学习/ NLP的模型。

## PySpark

我们先来回顾一下之前我们是如何在Hadoop上构建机器学习/ NLP模型，以及如何对这个Hadoop上ML模型进行评分的。在上一节中，我们较为深入地讨论了一下评分的第二个选项。与针对规模较小的数据集进行抽样不同的是，我们这次要对规模交大的数据集来进行评分，并使用PySpark库来逐步构建大型的机器学习模型。当然，我们在这里会再次使用到与之前相同模式的相同运行数据：

|  |  |  |
| --- | --- | --- |
| **ID** | **评论** | **类别** |
| UA0001 | I tried calling you, The service was not up to the mark | 1 |
| UA0002 | Can you please update my phone no | 0 |
| UA0003 | Really bad experience | 1 |
| UA0004 | I am looking for an iPhone | 0 |
| UA0005 | Can somebody help me with my password | 1 |
| UA0006 | Thanks for considering my request for | 0 |

通常，我们要考虑的是近10年来评价的组织模式。但现在我们要改用一个小型的样本，先为其创建分类模型，然后再用预训练模型对所有评论进行评分。下面，我们就通过一个例子来逐步演示一下这个过程，具体说明一下如何用PySpark库来构建文本分类模型。

首先来导入一些模块。我们从SparkContext开始，这更多的是一个配置性的模块，我们在这里可以通过更多的参数来提供信息，譬如如应用程序的名称等。

**>>> from pyspark import SparkContext**

**>>> sc = SparkContext(appName="comment\_classifcation")**

|  |
| --- |
| 如果想了解关于PySpark库的更多信息，您可以阅读一下下面这篇文章：  http://spark.apache.org/docs/0.7.3/api/pyspark/ pyspark.context.SparkContext-class.html。 |

接下来要读取这个由制表符分隔的文本文件。我们要将该文件读取到HDFS上， 文件的体积可能会很大（~Tb/Pb）：

>>> lines = sc.textFile("testcomments.txt")

现在，lines就成为了包含语料库中所有行的一个列表：

>>> parts = lines.map(lambda l: l.split("\t"))

>>> corpus = parts.map(lambda row: Row(id=row[0], comment=row[1], class=row[2]))

这部分代码的作用获取是字段列表，其依据是各行中分隔字段的“\ t”符。

下面我们要将不同RDD对象中的语料库分解成[ID, comment, class (0,1)]的形式：

>>> comment = corpus.map(lambda row: " " + row.comment)

>>> class\_var = corpus.map(lambda row:row.class)

一旦我们收集完了这些评论信息，接下来的处理过程就与*第6章“文本分类”*中所做的事非常类似了，我们在这里会用scikit库来执行标识化处理，散列化vectorizer以及用vectorizer来计算TF、IDF和tf-idf。

下面来看看如何创建标识化处理、词汇频率和反向文档频率，其代码如下：

>>> from pyspark.mllib.feature import HashingTF

>>> from pyspark.mllib.feature import IDF

# https://spark.apache.org/docs/1.2.0/mllib-feature-extraction.html

>>> comment\_tokenized = comment.map(lambda line: line.strip().split(" "))

>>> hashingTF = HashingTF(1000) # to select only 1000 features

>>> comment\_tf = hashingTF.transform(comment\_tokenized)

>>> comment\_idf = IDF().fit(comment\_tf)

>>> comment\_tfidf = comment\_idf.transform(comment\_tf)

接下来，我们要将该类与tfidf RDD对象进行合并：

>>> finaldata = class\_var.zip(comment\_tfidf)

然后我来做一个典型测试，进行训练取样：

>>> train, test = finaldata.randomSplit([0.8, 0.2], seed=0)

现在来执行主分类命令，该命令与scikit库非常相似。我们使用的是逻辑回归算法，这是一种使用度很广的分类器。pyspark.mllib中提供了各种可用的算法。

|  |
| --- |
| 如果想了解更多关于pyspark.mllib的信息，可以参考以下链接中的资料： https://spark. apache.org/docs/latest/api/python/pyspark.mllib.html |

下面是朴素贝叶斯分类器的示例：

>>> from pyspark.mllib.regression import LabeledPoint

>>> from pyspark.mllib.classification import NaiveBayes

>>> train\_rdd = train.map(lambda t: LabeledPoint(t[0], t[1]))

>>> test\_rdd = test.map(lambda t: LabeledPoint(t[0], t[1]))

>>> nb = NaiveBayes.train(train\_rdd,lambda = 1.0)

>>> nb\_output = test\_rdd.map(lambda point: (NB.predict(point.features), point.label))

>>> print nb\_output

nb\_output命令中包含了对于测试样本的最终预测。这里需要了解的最重要的一件事是，上面的样本只有不到50行的内容，而我们在代码中所构建的是一个符合行业标准的文本分类操作，可以处理PB级规模的训练样本。

## 本章小结

下面来对本章做个小结，本章的目标是介绍如何在大数据环境中使用我们目前所学到这些概念。在这里，我们具体介绍了如何在Hadoop中使用NLTK和scikit这些库。也讨论了如何评分一个机器学习模型或基于NLP的操作。

为了介绍这些内容，我们主要举了三个业界最常见的用例来做演示。只要理解了这些示例，我们就能用好NLTK、scikit和PySpark这三个库中的大部分函数了。

本章基本上算是对大数据环境下的NLP和文本挖掘所做的一个速食性的简介。这毕竟是当前最为热门的话题之一，我们在示例代码中所讨论到到的每个术语和工具本身都足以写成一本书。我在这里试图给读者提供某种黑客方法，以便对大数据这种规模下的文本挖掘问题做一个介绍。我期望并鼓励读者多去阅读一些跟大数据有关的技术资料，譬如Hadoop、Hive、Pig和Spark等，并自己去试着实现一下本章所介绍地这些示例。

1. 译者注：MapReduce是一种编程方式，主要用于针对大规模数据集的并行计算。 [↑](#footnote-ref-1)
2. 译者注：由于原文针对的是英文环境，所以这里的例句就不翻译了，下同。 [↑](#footnote-ref-2)