Spark mllib API- feature

作者：Zing
原文链接：https://ld246.com/article/1461139570018
来源网站：链滴

spark中大部分的向量转换采用训练（fit）-转化（transform）形式，因此会有对应的训练类和模型。
该模块主要包括了，标准化、归一化、分词、特征选择

##pyspark.mllib.feature.Normalizer(p=2.0)
使用Lp范式对样本进行归一化。
若1<=p<float('inf')，使用 sum(abs(vector)^p)^(1/p)范式。
若p=float('inf')，使用max(abs(vector))范式。

- transform(vector)
  - 参数：vector - 需要正则化的RDD
  - 返回： 正则化的向量RDD

##pyspark.mllib.feature.StandardScaler(withMean=False, withStd=True)
使用训练集的列统计信息，通过修改均值和范围进行标准化

- fit(dataset):StandardScalerModel

计算均值和方差，并以模型保存，以便后续使用。相当于训练模型。

---

pyspark.mllib.feature.StandardScalerModel(java_mod)

表示可以把特征转化为正态分布的StandardScaler模型

- setWithMean(withMean)

参数为boolean，决定是否使用均值

- setWithStd(withStd)

参数为boolean，决定是否使用std

- transform(vector)

对特征进行标准变换

##pyspark.mllib.feature.HashingTF(numFeatures=1048576)
使用hash建立起 项-频度 映射。

- numFeatures：向量维度
- indexOf(term)：返回指定项的索引
- transform(document)：将输入转化为项-频度向量

##pyspark.mllib.feature.IDF(minDocFreq=0)

IDF为逆向文件频率，公式如下：

idf = log((m + 1) / (d(t) + 1))

其中m为文件总数，d(t)为出现项t的文件数。

- 参数：minDocFreq

通过minDocFreq参数，可以利用IDF过滤掉一些在文档中出现次数过少的词。若设置为0，则返回TF-DF

- 方法：fit(dataset)

计算数据集的IDF

##pyspark.mllib.feature.IDFModel(java_model)

IDF模型

- IDF()：返回当前IDF向量
- transform(x)：将TF向量转化为TF-IDF向量

##pyspark.mllib.feature.Word2Vec

Word2Vec 创建了一个表示语料库中词语的的向量。算法首先从语料库中创建一个词汇表，然后创建应到词汇表中单词的向量。在自然语言处理和机器学习算法中，该向量可以直接使用。

我们使用skip-gram模型实现，并且使用分层softmax方法来训练模型。

- fit(data):使用data进行训练，计算向量
- setLearningRate(learningRate)：设置初始学习率
- setMinCount(minCount)：设置最少出现的token次数，默认5
- setNumIterations(numIterations)：设置迭代次数，默认1
- setNumPartitions(numPartitions)：设置分区个数，默认1
- setSeed(seed)：设置随机种子
- setVectorSize(vectorSize)：设置向量维度，默认100

##pyspark.mllib.feature.Word2VecModel(java_model)

Word2Vec fit得到的模型

- findSynonyms(word, num)：找到指定word的num个同义词
- getVectors()：返回代表向量的单词表
- transform(word)：将单词转化为向量

---

pyspark.mllib.feature.ChiSqSelector(numTopFeatures)

创建一个卡方向量选择器，用于特征选择

- 参数：numTopFeatures 保留的卡方较大的特征的数量。
- fit(data)：对LabeledPoint格式的RDD进行训练，返回ChiSqSelectorModel，这个类将输入数据化到降维的特征空间。

##pyspark.mllib.feature.ChiSqSelectorModel(java_model)

由ChiSqSelector训练得到的模型

- transform(vector)，对RDD进行转换，转化到降维的特征空间。

##pyspark.mllib.feature.ElementwiseProduct(scalingVector)

使用输入的scalingVector作为每一列的权值，对每一列进行扩展。

- transform(vector)：对向量进行Hadamard卷积。