In [46]:

```
from pyspark.sql import SQLContext, Row
from pyspark.ml.feature import CountVectorizer
from pyspark.mllib.clustering import LDA, LDAModel
from pyspark.mllib.linalg import Vector, Vectors
```

In [51]:

```
path = "/user/ncn251/cookbook_text1.zip"

def zip_extract(x):
    in_memory_data = io.BytesIO(x[1])
    file_obj = zipfile.ZipFile(in_memory_data, "r")
    files = [i for i in file_obj.namelist()]
    return [file_obj.open(file).read() for file in files]

zips=sc.binaryFiles(path,100)
zipData=sc.parallelize(zips.map(zip_extract).collect(),100)

data = zipData.zipWithIndex().map(lambda words: Row(idd=words[1],words=words[0].split("
")))
```

In [52]:

```
docDF = spark.createDataFrame(data)
Vector = CountVectorizer(inputCol="words", outputCol="vectors")
model = Vector.fit(docDF)
result = model.transform(docDF)
```

In [53]:

```
corpus = result.select("idd", "vectors").rdd.map(lambda x: [x[0],Vectors.fromML(x[1])])
.cache()
corpus
```

Out[53]:

```
PythonRDD[793] at RDD at PythonRDD.scala:48
```

In [54]:

```
# Cluster the documents into three topics using LDA
ldaModel = LDA.train(corpus, k=3,maxIterations=100,optimizer='online')
topics = ldaModel.topicsMatrix()
vocabArray = model.vocabulary
```

In [55]:

```
wordNumbers = 100 # number of words per topic
topicIndices = sc.parallelize(ldaModel.describeTopics(maxTermsPerTopic = wordNumbers))
```

In [56]:

```
def topic_render(topic):  # specify vector id of words to actual words
    terms = topic[0]
    result = []
    for i in range(wordNumbers):
        term = vocabArray[terms[i]]
        result.append(term)
    return result
```

In [57]:

```
topics_final = topicIndices.map(lambda topic: topic_render(topic)).collect()
```

In [61]:

```
for topic in range(len(topics_final)):
    print ("Topic" + str(topic) + ":")
    print(topics_final[topic])
#     for term in topics_final[topic]:
#         print (term)
    print ('\n')
```

```
Topic0:
['.', 'he', 'his', 'was', 'I', 'had', 'her', 'she', 'my', 'said', 'were',
'', 'old', 'their', 'him', 'S', 'the', 'who', 'He', 'man', 'me', 'that',
'how', 'young', 'Zuñi', 'our', 'shall', 'we', 'to', 'came', 'did', 'Mary',
'what', 'thought', 'you', 'would', 'see', 'went', 'could', 'thou', 'peopl
e', 'him,', 'woman', 'but', 'down', 'told', 'tell', '"I', 'yet', 'She', 'h
ouse', 'of', 'toward', 'women', 'men', 'he,', 'thy', 'began', 'poor', 'anc
ient', 'might', 'And', 'youth', 'us', 'took', 'and', 'for', 'grew', 'hear
d', 'looked', 'like', 'saw', 'sat', 'boy', 'am', 'at', 'great', 'go', 'cor
n', 'man,', 'know', 'himself', 'girl', 'forth', 'father', 'gave', 'ever',
'knew', 'asked', 'deer', 'home', 'hunter', 'So', 'ye', 'Time.', '"', 'm
e,', 'they', 'brought', 'by']


Topic1:
['', ' ', '1', '-', '"', '&#224;', 'la', '--', 'de', '2', 'AND', 'N
o.', '.....', 'OF', 'or', 'of,', 'for', '3', 'A', 'Cream', 'To', 'Fruit',
'TO', 'WITH', '4', 'FOR', '1/2', 'Sauce', 'au', 'with', 'See', 'LA', 'Brea
d', 'Baked', 'OR', 'Stewed', 'Boiled', 'Roast', 'Soup', 'White', 'MRS.',
'Coffee', 'THE', 'Apple', 'aux', 'Rice', '6', 'NO.', 'Chicken', 'Potato',
'SAUCE', '&#32;', 'Mrs.', 'Corn', 'Beef', '1/4', 'Sauce,', 'lb.', 'Lemon',
'Tomato', 'French', 'Fried', 'Pudding.', 'M', 'Graham', 'Green', 'Puddin
g,', '(', 'BREAKFAST', '5', 'salad', '8', 'Fresh', 'en', 'Sweet', 'DINNE
R', 'Broiled', 'Fish', '30', 'Soup,', 'et', 'Sauce.', '20', 'Eggs', '10',
'Orange', '&#192;', 'potatoes', 'Pudding', 'Cake', 'CREAM', '..', 'see',
'CAKE.', 'Egg', ',', 'IN', 'Salt', 'Salad', 'sauce']


Topic2:
['', 'the', 'and', 'of', 'a', 'in', 'to', 'with', 'it', 'is', 'or', 'be',
'as', 'for', 'one', 'them', 'into', 'on', 'are', 'two', 'then', 'add', 'fr
om', 'put', 'an', 'that', 'over', 'not', 'half', 'will', 'little', 'whic
h', 'when', 'by', 'this', 'The', 'all', 'very', 'until', 'water', 'they',
'at', 'should', 'three', 'cut', 'may', 'have', 'but', 'some', 'cup', 'smal
l', 'pound', 'if', 'well', 'you', 'cold', 'butter', 'hot', 'water,', 'le
t', 'boil', 'up', 'out', 'pint', 'boiling', '1', 'sugar', 'more', 'pour',
'sugar,', 'salt', 'four', 'When', 'each', 'make', 'about', 'large', 'mad
e', 'so', 'take', 'butter,', 'white', 'before', 'has', 'salt,', 'stir', 'c
an', 'cover', 'good', 'off', 'flour', 'eggs', 'It', 'same', 'their', 'thro
ugh', 'Put', 'teaspoonful', 'than', 'place']
```