In [10]:

```
#function that removes punctuations
def removePunctuation(text):
    return re.sub(r'[^a-z0-9\s]','',text.lower().strip())
#function to count the word occurence
def lenCheck(word):
    if len(word)>0:
        return word
import os.path
import re
from operator import add
#path where the data file is stored
baseDir=os.path.join("PracticeData")
inputDir=os.path.join("big.txt")
fileName=os.path.join(baseDir,inputDir)
#creating a base RDD and removing punctuations
fileRDD=(sc
        .textFile(fileName,8)
        .map(removePunctuation))
#splitting the words with space and removing the empty lines
fileWordsRDD=fileRDD.flatMap(lambda line:line.split())
#remove empty spaces
fileWordsRDD=fileWordsRDD.map(lenCheck)
#count the number of words. (key,count)
fileCountRDD=fileWordsRDD.map(lambda x:(x,1)).reduceByKey(add)
# to display to 15 words and their counts
top15words=fileCountRDD.takeOrdered(15,key=lambda (w,c):-c)
#print top15 words in word,count format

print '\n'.join(map(lambda (w,c):'{0}:  {1}'.format(w,c),top15words))
```

```
the:  79389
of:  39996
and:  38092
to:  28611
in:  21776
a:  20871
he:  12182
that:  12090
was:  11368
it:  10163
his:  10014
is:  9731
with:  9705
as:  7948
had:  7364
```