In [1]:

```
import findspark
findspark.init()
import pyspark
```

In [20]:

```
def mapwords(sentence) :
    liste = []
    for word in sentence :
        liste.append((word,1))
    return liste
```

In [53]:

```
def removespecialchars (word) :
    charstoremove = ['[',']','#','.','_',':','?','!',',',';','“','”','\n']
    for ch in word:
        if ch in charstoremove:
            word = word.replace(ch,'')
    return word
```

In [3]:

```
sc = pyspark.SparkContext(appName="wordcount")
```

In [ ]:

```
#Read file in utf-8 using use_unicode=False (default kept incoding is utf-8 in this case)
rdd = sc.textFile("C:/Users/EASYFRONT/Desktop/bigdata/seance01/BigWordCount.txt", use_unicode="False")
rdd1 = rdd
#rdd = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/Wordcount.txt")
rdd = rdd.map (lambda x : x.split(" "))
rdd = rdd.flatMap (lambda x : mapwords(x))
rdd = rdd.reduceByKey (lambda x,y : x+y)
rdd.collect()
```

In [ ]:

```
#Jane Austen Example
rdd1= sc.textFile("C:/Users/EASYFRONT/Desktop/bigdata/seance01/wordcount.txt", use_unicode="False")
todrop = rdd1.take(30)
todrop
rdd1 = rdd1.filter (lambda x : x not in todrop)
rdd1 = rdd1.flatMap(lambda x : x.split(" "))
rdd1 = rdd1.map(lambda x : removespecialchars(x))
rdd1 = rdd1.map(lambda x : x.lower())
rdd1 = rdd1.filter(lambda x : x!='')
nbwords = rdd1.count()
rdd1 = rdd1.map(lambda x : (x,1))
rdd1 = rdd1.reduceByKey (lambda x,y : x+y)
frequency = rdd1.map(lambda x : (x[0], x[1]/nbwords))
#rdd1.collect()
#nbwords
frequency = frequency.sortBy(lambda x : x[1],0)
frequency.collect()
```