**Patterns for pattern matching.**

pattern1='((\d\d\d:\d\d\d{1}))'  
pattern2='(\?|\.|,|;|!|:)\*(\w+(\?|\.|,|;|!|:|\')\*(\w+))(\?|\.|,|;|!|:)\*'  
pattern3='(\w+)(\'s)'

**Precompiling patterns to increase efficiency**

RE for date filtering

regdate=re.compile(pattern1)

RE for Special character filtering (removing from start and end)

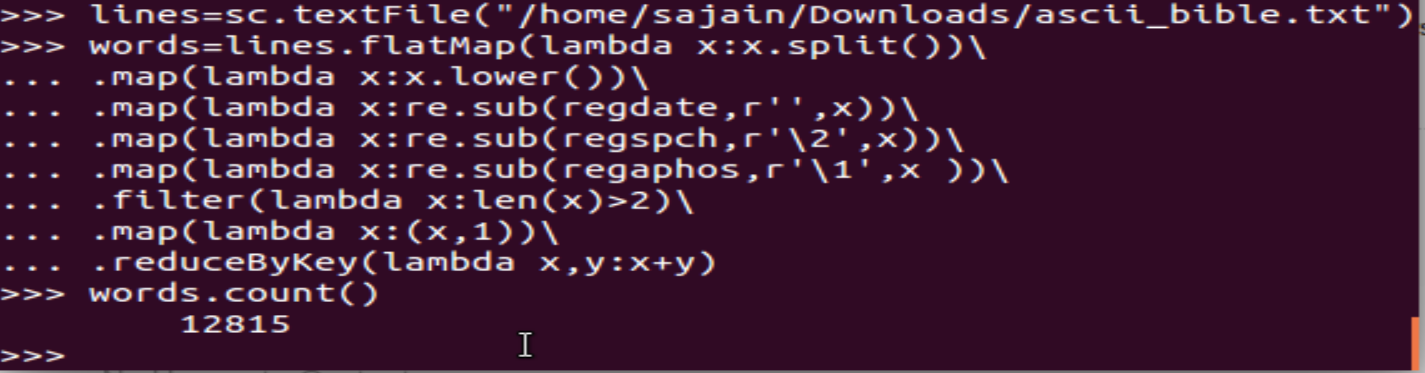
regspch=re.compile(pattern2)

RE for apostrophe handling

regaphos=re.compile(pattern3)

**code**

lines=sc.textFile("/home/sajain/Downloads/ascii\_bible.txt")  
words=lines.flatMap(lambda x:x.split())\  
.map(lambda x:x.lower())\  
.map(lambda x:re.sub(regdate,r'',x))\  
.map(lambda x:re.sub(regspch,r'\2',x))\  
.map(lambda x:re.sub(regaphos,r'\1',x ))\

.filter(lambda x:len(x)>2)\  
.map(lambda x:(x,1))\  
.reduceByKey(lambda x,y:x+y)  
**Output :**  


words.count()

**output 1:**

12815

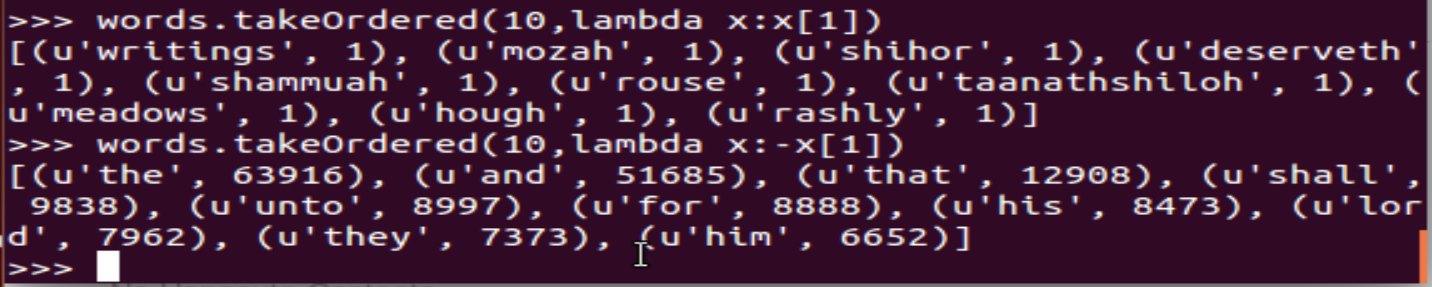
To show top 10 in the word count

words.takeOrdered(10,lambda x:x[1])

To show bottom 10 in the word count

words.takeOrdered(10,lambda x:-x[1])

**output 2 & 3:**



words.saveAsTextFile(“output”)

ls –l output

**output 4:**

