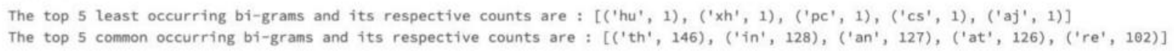
Programming Assignment 1

```
from operator import add

from string import punctuation

data = sc.textFile("/FileStore/tables/Class_6___Text_File_for_Letter_Pairs-
5.txt").map(lambda x: x.replace(',','').replace('”', '').replace('-', '').replace('$',
'').replace('.', '').replace(')', '').replace('(','').replace('$', '').replace(')',
'').replace('-','').replace('“', '').replace('€', '').replace('’', '').replace('‘',
'').replace('“', '').replace('0', '').replace('1', '').replace('2', '').replace('3',
'').replace('4', '').replace('5', '').replace('6', '').replace('7', '').replace('8',
'').replace('9', '').lower())


bigram = data.flatMap(lambda x:x.lower().strip(punctuation).split("
")).flatMap (lambda s:[((s[i] + s[i+1]),1) for i in range (0, len(s)-
1)]).reduceByKey(lambda x,y : x+y)


#top n least occurring bi-grams

output_min = bigram.sortBy(lambda x: x[1])

print("The top {0} least occurring bi-grams and its respective counts are :
{1}". format(5, output_min.take(5)))


#top n common occurring bi-grams

output_max = bigram.sortBy(lambda x: -x[1])

print("The top {0} common occurring bi-grams and its respective counts are :
{1}".format(5, output_max.take(5)))
```

Output



```
The top 5 least occurring bi-grams and its respective counts are : [('hu', 1), ('xh', 1), ('pc', 1), ('cs', 1), ('aj', 1)]
The top 5 common occurring bi-grams and its respective counts are : [('th', 146), ('in', 128), ('an', 127), ('at', 126), ('re', 102)]
```