

```

1 !pip install pyspark

1 import pyspark
2 pyspark
3 from pyspark.context import SparkContext, SparkConf

1 conf = SparkConf().setAppName("assignment1").setMaster("local")
2 sc = SparkContext(conf=conf)

1 data = sc.textFile("/content/drive/MyDrive/Spring 2022/BIA 678 big data tech/program")
2 data_1 = data.map(lambda z: "".join(z))
3 data_2 = data_1.map(lambda z: z.replace(" " , ""))
4 data_3 = data_2.map(lambda z: z.replace(".", ""))
5 data_final = data_3.map(lambda z: z.lower())

1 bigrams = data_final.flatMap(lambda s: [(s[i],s[i+1]),1] for i in range (0, len(s)-1))
2 frequency = bigrams.reduceByKey(lambda x,y: x+y)
3 reversed_frequency = frequency.map(lambda reverse:(reverse[1],reverse[0]))

1 print('Five most frequent Bigrams: \n')
2 reversed_frequency.sortByKey(True).take(5)

↪ Five most frequent Bigrams:

[(1, ('a', 'o')),
 (1, ('y', 'l')),
 (1, ('l', 'c')),
 (1, ('h', 'u')),
 (1, ('s', '2'))]

1 print('\nFive Least frequent Bigrams: \n')
2 reversed_frequency.sortByKey(False).take(5)

```

Five Least frequent Bigrams:

```

[(147, ('t', 'h')),
 (134, ('a', 't')),
 (128, ('i', 'n')),
 (127, ('a', 'n')),
 (106, ('r', 'e'))]

```

```

1 five_most_freq = {''.join(k): v for k, v in sorted(frequency.collect(), key=lambda
2 print('Five most frequent Bigrams: \n')
3 print(five_most_freq)

```

Five most frequent Bigrams:

```
{'th': 147, 'at': 134, 'in': 128, 'an': 127, 're': 106}
```

```
1 five_least_freq = {''.join(k): v for k, v in sorted(frequency.collect(), key=lambda  
2 print('Five Least frequent Bigrams: \n')  
3 print(five_least_freq)
```

Five Least frequent Bigrams:

```
{'ao': 1, 'yl': 1, 'lc': 1, 'hu': 1, 's2': 1}
```

1

