```
from pyspark.sql import SparkSession
import re
from pyspark.dbutils import DBUtils
spark  = SparkSession.builder.appName('bigram tester').getOrCreate()
sc= spark.sparkContext

dbutils = DbUtils(spark)
dbutils.fs.ls('dbfs:/FileStore/tables')

words = sc.textFile('dbfs:/FileStore/tables/letter_pair.txt')
words.collect()

bigram_pairs = words.map(lambda x: x.lower()).flatMap(lambda x: re.sub(r"[^a-zA-
Z]","",x)).flatMap(lambda s: [((s[i : i+2], 1) for i in range (0, len(s)-1)]).filter(lambda s:
len(s[0]==2)
                .filter(lambda s: not str(s[0]).isnumeric())
bigram_pairs.collect()

bigram_count = bigram_pairs.reduceByKey(lambda x, y : x+y).maps(lambda x: (x[1],
x[0])).sortByKey(False)
print(bigram_count.take(5))

print(bigram_count.take(bigram_count.count())[bigram_count.count()-5:])
```

top 5 frequent pairs

```
1 sorted(bigram_dict.items(), key=lambda x: x[1], reverse=True)[:5]
```

```
[('th', 136), ('at', 126), ('an', 126), ('in', 120), ('re', 102)]
```

top 5 least frequent pairs

```
1 print(bigram_count.take(bigram_count.count())[bigram_count.count()-5:])
```

```
[('hu', 1), ('xh', 1), ('yp', 1), ('pc', 1), ('cs', 1)]
```

```
from pyspark.sql import SparkSession
import re
from pyspark.dbutils import DBUtils
spark  = SparkSession.builder.appName('bigram tester').getOrCreate()
sc= spark.sparkContext

dbutils = DbUtils(spark)
dbutils.fs.ls('dbfs:/FileStore/tables')

words = sc.textFile('dbfs:/FileStore/tables/letter_pair.txt')
words.collect()

bigram_pairs = words.map(lambda x: x.lower()).flatMap(lambda x: re.sub(r"[^a-zA-Z]","",x)).flatMap(lambda s: [((s[i
                        .filter(lambda s: not str(s[0]).isnumeric())
bigram_pairs.collect()

bigram_count = bigram_pairs.reduceByKey(lambda x, y : x+y).maps(lambda x: (x[1], x[0])).sortByKey(False)
print(bigram_count.take(5))

print(bigram_count.take(bigram_count.count())[bigram_count.count()-5:])
```