

2015125

Lab - 10

write a simple streaming prog
in spark to receive text data
streams on a particular port,
perform basic cleaning and print
the output.

```
import org.apache.spark._  
import org.apache.spark.streaming._
```

```
val ssc = new StreamingContext(  
    sc, Seconds(5))
```

```
val line = ssc.socketTextStream(  
    "localhost", 9999)
```

```
val cleanLines = lines.map {  
    line =>
```

```
    val low = line.toLowerCase.trim  
    replaceAll("\\s+", " ")
```

```
    val noPunct = low.replaceAll(  
        "[^a-zA-Z\\s]", "")
```

```
    val tokens = noPunct.split(" ")  
    .filter(token => !stopWords  
        contains(token))
```

```
    val lemmalized = tokens.map {  
        word => word }
```


lemmatized.mkString(" ")

ssc.start()

ssc.awaitTermination()

Output

stream

Hello · World · the weather is cold

⇒ ~~Hello World the weather cold~~