# Assignment1

```
//Importing library
import com.johnsnowlabs.nlp.base._
import com.johnsnowlabs.nlp.annotator._
import com.johnsnowlabs.nlp.pretrained.PretrainedPipeline
import com.johnsnowlabs.nlp.SparkNLP
import org.apache.spark.sql.functions._
import org.apache.spark.sql.Row
```

Cmd 2

```
//Reading the file and filtering
val txtfile = sc.textFile("/FileStore/tables/largesttt56613_0-1.txt")
val dataDF = textfile.filter(x=>x.length>0).zipWithIndex.toDF("text","_id").select("_id","text") //Filtering
dataDF.take(10)
```

```
txtfile: org.apache.spark.rdd.RDD[String] = /FileStore/tables/largesttt56613_0-1.txt MapPartitionsRDD[143] at textFile at command-2187602793455931:1
```

Command took 0.34 seconds -- by kxn190007@utdallas.edu at 9/23/2020, 5:57:17 PM on UTD1

Cmd 3

```
//Using Library
val pipeline = PretrainedPipeline("recognize_entities_dl", "en")
val predictions = pipeline.transform(dataDF)
pipeline.transform(dataDF).select("entities.result")
val output = pipeline.transform(dataDF).select(explode($"entities.result"))
output.show()
```

- (1) Spark Jobs
- predictions: org.apache.spark.sql.DataFrame = [_id: long, text: string ... 6 more fields]
- output: org.apache.spark.sql.DataFrame = [col: string]

```
+--------------------+
|                 col|
+--------------------+
|Project Gutenberg...|
|             English|
|         Andrew Lang|
|       United States|
|Project Gutenberg...|
|               eBook|
|       United States|
```