

## Lab 10

**Write a simple streaming program in Spark to receive text data streams on a particular port, perform basic text cleaning (like white space removal, stop words removal, lemmatization, etc.), and print the cleaned text on the screen. (Open Ended Question).**

```
# Install NLTK and download required data
(run once) !pip install nltk

import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, lower,
regexp_replace, split, explode, udf from pyspark.sql.types
import ArrayType, StringType from pyspark.ml.feature
import StopWordsRemover from nltk.stem import
WordNetLemmatizer

#      Initialize      SparkSession      spark      =
SparkSession.builder.appName("TextProcessing").getO
rCreate()

# Define your input lines
lines = [
    "Hello, I hate you.",
    "I hate that I love you.",
    "Don't want to, but I can't put",
    "nobody else above you."
]

# Create DataFrame from lines
df = spark.createDataFrame(lines, "string").toDF("value")

# Step 1: Lowercase and remove punctuation df_clean =
df.select(regexp_replace(lower(col("value")), "[^a-zA-Z\\s]",
""), alias("cleaned"))

# Step 2: Tokenize the cleaned text df_tokens =
df_clean.select(split(col("cleaned"),
"\\s+").alias("tokens"))

# Step 3: Remove stop words
```

```
remover      =      StopWordsRemover(inputCol="tokens",
outputCol="filtered")      df_filtered      =
remover.transform(df_tokens)
```

# Step 4: Lemmatization using NLTK WordNetLemmatizer  
with UDF lemmatizer = WordNetLemmatizer()

```
def lemmatize_words(words):
    return [lemmatizer.lemmatize(word) for word in words]
```

```
lemmatize_udf = udf(lemmatize_words, ArrayType(StringType()))
df_lemmatized = df_filtered.withColumn("lemmatized", lemmatize_udf(col("filtered")))
df_lemmatized.select(explode(col("lemmatized")).alias("word")).show(truncate=False)
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.2.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.5.0)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
+-----+
|word |
+-----+
|hello |
|hate  |
|hate  |
|love  |
|dont  |
|want  |
|cant  |
|put   |
|nobody|
|else  |
+-----+
```