Tokenizer Example in Apache openNLP using Java

Tokenizer Example in Apache openNLP

In this openNLP Tutorial, we shall look into Tokenizer Example in Apache openNLP. Also, a little understanding of the tokenizaion process.

What is tokenization ?

Tokenization is a process of segmenting strings into smaller parts called tokens(say sub-strings). These tokens are usually words, punctuation marks, sequence of digits, and like. An example is shown in the following table :

Input to Tokenizer	John is 26 years old.
Output of Tokenizer	John | is | 26 | years | old | .

Tokenization in OpenNLP

Tokenizer API in OpenNLP provides following three ways for tokenization :

- TokenizerME class loaded with a token model
- WhitespaceTokenizer
- SimpleTokenizer

Note : OpenNLP version used is 1.7.2.

Please observe the differences in the output from these three ways of tokenization in the examples provided below.

TokenizerME class loaded with a token model

- **Step 1** : Read the pretrained model into a stream.

```
InputStream mod
```

```
InputStream modelIn = new FileInputStream("en-token.bin");
```

- **Step 2** : Read the stream to a Tokenizer model.

```
TokenizerModel model
```

```
TokenizerModel model = new TokenizerModel(modelIn);
```

- **Step 3** : Initialize the tokenizer with the model.

```
TokenizerME tokenizer
= new
```

```
TokenizerME tokenizer = new TokenizerME(model);
```

- **Step 4** : Use TokenizerME.tokenize() method to extract the tokens to a String Array.

```
String tokens[] =
tokenizer.tokenize("Joh
```

```
String tokens[] = tokenizer.tokenize("John is 26 years old.");
```

- **Step 5** : Use TokenizerME.getTokenProbabilities() to get the probabilities for the segments to be tokens.

```
double tokenProbs[] =
tokenizer.getTokenPro
```

```
double tokenProbs[] = tokenizer.getTokenProbabilities();
```

- **Step 6** : Finally, print the results.

Everything put together, is the below below program :

TokenizerModelExample.java

```
import
java.io.FileInputStream
```

```
import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStream;

import opennlp.tools.tokenize.TokenizerME;
import opennlp.tools.tokenize.TokenizerModel;

/**
* www.tutorialkart.com
* Tokenizer Example in Apache openNLP using TokenizerME class loaded with pre-trained token model
*/
public class TokenizerModelExample {

    public static void main(String[] args) {
        InputStream modelIn = null;

        try {
            modelIn = new FileInputStream("en-token.bin");
            TokenizerModel model = new TokenizerModel(modelIn);
            TokenizerME tokenizer = new TokenizerME(model);
            String tokens[] = tokenizer.tokenize("John is 26 years old.");
            double tokenProbs[] = tokenizer.getTokenProbabilities();

            System.out.println("Token\t: Probability\n-------------------------------");
            for(int i=0;i<tokens.length;i++){
                System.out.println(tokens[i]+"\t: "+tokenProbs[i]);
            }
        }
        catch (IOException e) {
            e.printStackTrace();
        }
        finally {
            if (modelIn != null) {
                try {
                    modelIn.close();
                }
                catch (IOException e) {
                }
            }
        }
    }
}
```

When the above program is run, the output to the console is as shown below :

Program Output

```
Token : Probability
```

```
Token : Probability
--------------------------------
John : 1.0
is : 1.0
26 : 1.0
years : 1.0
old : 0.9954218897531331
. : 1.0
```

WhitespaceTokenizer

Following is the example to demostrate WhitespaceTokenizer of OpenNLP Tokenizer API

WhiteSpaceTokenizerExample.java

```
import
opennlp.tools.tokenize
```

```
import opennlp.tools.tokenize.Tokenizer;
import opennlp.tools.tokenize.WhitespaceTokenizer;

/**
* www.tutorialkart.com
* Tokenizer Example in Apache openNLP using WhitespaceTokenizer
*/
public class WhiteSpaceTokenizerExample {

    public static void main(String[] args) {
        Tokenizer tokenizer = WhitespaceTokenizer.INSTANCE;
        String tokens[] = tokenizer.tokenize("John is 26 years old.");

        System.out.println("Token\n----------------");
        for(int i=0;i<tokens.length;i++){
            System.out.println(tokens[i]);
        }
    }
}
```

When the above program is run, the output to the console is as shown below :

Program Output

```
Token
----------------
```

```
Token
----------------
John
is
26
years
old.
```

SimpleTokenizer

Following is the example to demostrateSimpleTokenizer of OpenNLP Tokenizer API

SimpleTokenizerExample.java

```
import
opennlp.tools.tokenize
```

```
import opennlp.tools.tokenize.SimpleTokenizer;
import opennlp.tools.tokenize.Tokenizer;

/**
* www.tutorialkart.com
* Tokenizer Example in Apache openNLP using SimpleTokenizer
*/
public class SimpleTokenizerExample {

    public static void main(String[] args) {
        Tokenizer tokenizer = SimpleTokenizer.INSTANCE;
        String tokens[] = tokenizer.tokenize("John is 26 years old.");

        System.out.println("Token\n----------------");
        for(int i=0;i<tokens.length;i++){
            System.out.println(tokens[i]);
        }
    }
}
```

When the above program is run, the output to the console is as shown below :

Program Output

```
Token
```

```
Token
----------
John
is
26
years
old
.
```

Conclusion :

In this Apache OpenNLP Tutorial, we have seen different ways of tokenization the OpenNLP Tokenizer API provides.

Following are some of the other examples of openNLP :

- Named Entity Extraction
- Parts-Of-Speech Tagging
- Sentence Detection

Learn OpenNLP

- OpenNLP Tutorial
- Setup Java Project with OpenNLP in Eclipse
- OpenNLP Models

Detection / Extraction using Java API

- Tokenizer Example
- Sentence Detection Example
- Parts-Of-Speech Tagger Example
- Chunker Example
- Lemmatizer Example
- Named Entity Extraction Example

Training using Java API

- Sentence Detection Model Training
- Name Entity Finder Model Training
- Document Categorizer Training - Maximum Entropy
- Document Categorizer Training - Naive Bayes
- Document Categorizer with N-gram features used
- Language Detector Training Example

Command Line Tools

- Setup and start using Command Line Tools

Useful Resources

- How to Learn Programming