Lemmatizer Example in Apache OpenNLP

Lemmatizer is a Natural Language Processing tool that aims to remove any changes in form of the word like tense, gender, mood, etc. and return dictionary or base form of word.

In Apache OpenNLP, Lemmatizer returns base or dictionary form of the word (usually called lemma) when it is provided with word and its Parts-Of-Speech tag. For a given word, there could exist many lemmas, but given the Parts-Of-Speech tag also, the number could be narrowed down to almost one, and the one is the more accurate as the context to the word is provided in the form of postag.

In Apache OpenNLP there are two methods to do Lemmatization.

- Statistical Lemmatization
- Dictionary based Lemmatization

Statistical Lemmatizer needs a lemmatizer model(that is built from training data) for finding the lemma of a given word, while the Dictionary based Lemmatizer needs a dictionary(which contains all possible and valid combinations of {word, postag and the corresponding lemma}) .

Input to the Lemmatizer is the set of tokens and corresponding postags. So, to find lemmas for words in a sentence, the prior task is : sentence has to be tokenized using a Tokenizer and then pos tagged using a POS Tagger.

Dictionary Lemmatizer Example in Apache OpenNLP

You may download the dictionary from here[https://raw.githubusercontent.com/richardwilly98/elasticsearch-opennlp-auto-tagging/master/src/main/resources/models/en-lemmatizer.dict]. And en-pos-maxent.bin from here[http://opennlp.sourceforge.net/models-1.5/].

DictionaryLemmatizerExample.java

```
import
```

```
import opennlp.tools.langdetect.*;
import opennlp.tools.lemmatizer.DictionaryLemmatizer;
import opennlp.tools.postag.POSModel;
import opennlp.tools.postag.POSTaggerME;

import java.io.*;

/**
* Dictionary Lemmatizer Example in Apache OpenNLP
*/
public class DictionaryLemmatizerExample {

    public static void main(String[] args){
```

```
        try{
            // test sentence
            String[] tokens = new String[]{"Most", "large", "cities", "in", "the", "US", "had",
                    "morning", "and", "afternoon", "newspapers", "."};

            // Parts-Of-Speech Tagging
            // reading parts-of-speech model to a stream
            InputStream posModelIn = new FileInputStream("models"+File.separator+"en-pos-maxent.bin");
            // loading the parts-of-speech model from stream
            POSModel posModel = new POSModel(posModelIn);
            // initializing the parts-of-speech tagger with model
            POSTaggerME posTagger = new POSTaggerME(posModel);
            // Tagger tagging the tokens
            String tags[] = posTagger.tag(tokens);

            // loading the dictionary to input stream
            InputStream dictLemmatizer = new FileInputStream("dictionary"+File.separator+"en-lemmatizer.txt");
            // loading the lemmatizer with dictionary
            DictionaryLemmatizer lemmatizer = new DictionaryLemmatizer(dictLemmatizer);

            // finding the lemmas
            String[] lemmas = lemmatizer.lemmatize(tokens, tags);

            // printing the results
            System.out.println("\nPrinting lemmas for the given sentence...");
            System.out.println("WORD -POSTAG : LEMMA");
            for(int i=0;i< tokens.length;i++){
                System.out.println(tokens[i]+" -"+tags[i]+" : "+lemmas[i]);
            }

        } catch (FileNotFoundException e){
            e.printStackTrace();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

Output :

Program Output

```
/usr/lib/jvm/default-
java/bin/java -
```

```
/usr/lib/jvm/default-java/bin/java -javaagent:/media/arjun/0AB650F1B650DF2F/SOFTs/ubuntu/idea-IC-171.4249.39/lib/idea_rt.
DictionaryLemmatizerExample

Printing lemmas for the given sentence...
WORD -POSTAG : LEMMA
Most -JJS : much
large -JJ : large
cities -NNS : city
in -IN : in
the -DT : the
US -NNP : O
had -VBD : have
morning -NN : O
and -CC : and
afternoon -NN : O
newspapers -NNS : newspaper
. -. : O

Process finished with exit code 0
```

Note : If a combination of the word and postag is not found in the dictionary, the lemma is returned as '0' (like zero findings). In the above example the combinations US_NNP, morning_NN, afternoon_NN and ._. are not found in the dictionary, hence the corresponding lemmas are '0'.

Conclusion :

We have learnt what is lemmatization and how to implement it, with the help of Lemmatizer Example in Apache OpenNLP.

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