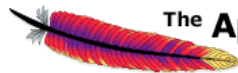


Natural Language Processing



The **Apache Software Foundation**
<http://www.apache.org/>

Open NLP (<http://opennlp.apache.org/>)

- Java library for processing natural language text
 - Based on Machine Learning tools
 - maximum entropy, perceptron
 - Includes pre-built models for some languages and annotated text resources
 - Is work in progress....
- Supported NLP tasks
 - tokenization
 - sentence segmentation
 - part-of-speech tagging
 - named entity extraction
 - chunking
 - parsing
 - coreference resolution (experimental)

Library structure

- The library provides components to approach specific NLP tasks
 - The components can be combined to build a NLP processing pipeline
 - Each component interface in general has methods for
 - execute the NLP processing task on a given input text stream
 - train a model for the NLP task from examples
 - evaluate a model on test data
 - The component functionalities can be accessed through a Java API or a command line interface (CLI)
 - read the model from file
 - instantiate the model
 - execute the processing task

```
SomeModel model = new SomeModel(  
    new FileInputStream("lang-model-name.bin"));
```

```
ToolName toolName = new ToolName(model);
```

```
String output[] = toolName.executeTask(  
    "This is a sample text.");
```

CLI command

- The **opennlp** script allows to exploit the available modules

```
OpenNLP 1.5.3. Usage: opennlp TOOL
where TOOL is one of:
  Doccat                      learnable document categorizer
  DoccatTrainer               trainer for the learnable document categorizer
  DoccatConverter             converts leipzig data format to native OpenNLP
  format
  DictionaryBuilder           builds a new dictionary
  SimpleTokenizer             character class tokenizer
  TokenizerME                 learnable tokenizer
  TokenizerTrainer            trainer for the learnable tokenizer
  TokenizerMEEvaluator        evaluator for the learnable tokenizer
  TokenizerCrossValidator     K-fold cross validator for the learnable tokenizer
  .....
  ChunkerME                   learnable chunker
  ChunkerTrainerME            trainer for the learnable chunker
  .....
  Coreferencer                learnable noun phrase coreferencer
  CoreferencerTrainer
  ....
```

Name Finder

- Detection of Named Entities and numbers in text
 - A trainable model is exploited to detect the entities
 - The model depends on the language and on the entity type
 - A set of pre-trained models is available in the OpenNLP library
 - en-ner-date, en-ner-location, en-ner-money, en-ner-organization, en-ner-percentage, en-ner-person, en-ner-time
 - The processing needs is performed on the tokenized text

Person finder

```
> openNLP/bin/opennlp TokenNameFinder models/en-ner-person.bin
```

```
Loading Token Name Finder model ... done (1.585s)
```

```
Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov . 29 .
```

```
<START:person> Pierre Vinken <END> , 61 years old , will join the board as a nonexecutive  
director Nov . 29 .
```

Name types

- The model depends on a specific type

> `openNLP/bin/opennlp TokenNameFinder models/en-ner-date.bin`

Loading Token Name Finder model ... done (1.516s)

Pierre Vinken , 61 years old , will join the board as a nonexecutive director November 29 .

Pierre Vinken , 61 years old , will join the board as a nonexecutive director **<START:date>** November 29 **<END>** .

Date finder

- The available model does not work with the abbreviation Nov.

➤ `openNLP/bin/opennlp TokenNameFinder models/en-ner-organization.bin`

Loading Token Name Finder model ... done (1.606s)

The UN was founded in 1945 after World War II to replace the League of Nations , to stop wars between countries , and to provide a platform for dialogue .

The **<START:organization>** UN **<END>** was founded in 1945 after World War II to replace the **<START:organization>** League of Nations **<END>** , to stop wars between countries , and to provide a platform for dialogue .

Organization finder

Chunker

- Splits the text into syntactically correlated groups of words
 - noun groups, verb groups,...
 - the internal structure of a group is not explained
 - the group role in the sentence is not determined
 - the input is a PoS tagged text

```
➤ opennlp ChunkerME models/en-chunker.bin < example-POS.txt
```

Loading Chunker model ... done (1.058s)

[NP Pierre_NNP Vinken_NNP] ,_, [NP 61_CD years_NNS] [ADJP old_JJ] ,_, [VP will_MD join_VB] [NP the_DT board_NN] [PP as_IN] [NP a_DT nonexecutive_JJ director_NN] [NP Nov._NNP 29_CD] ._.

[NP Mr._NNP Vinken_NNP] [VP is_VBZ] [NP chairman_NN] [PP of_IN] [NP Elsevier_NNP N.V._NNP] ,_, [NP the_DT Dutch_JJ publishing_NN group_NN] ._.

[NP Rudolph_NNP Agnew_NNP] ,_, [NP 55_CD years_NNS] [ADJP old_JJ] and_CC [ADVP former_JJ] [NP chairman_NN] [PP of_IN] [NP Consolidated_NNP Gold_NNP Fields_NNP PLC_NNP] ,_, [VP was_VBD named_VBN] [NP a_DT director_NN] [PP of_IN] [NP this_DT British_JJ industrial_JJ conglomerate_NN] ._.