Introduction to the **openNLP** Package

Ingo Feinerer

April 14, 2008

Abstract

The **openNLP** package.

Introduction

The **openNLP** package provides a R interface to openNLP[1].

Loading the Package

The package is loaded via

```
> library("openNLP")
```

Models

It is highly recommended to use the **openNLPmodels** package to provide the models necessary for full **openNLP** functionality.

```
> library("openNLPmodels")
```

This package provides default models both for English (`en`) and Spanish (`es`).

Part-of-speech Tagging

```
> sentence <- "This is a short sentence consisting of
+              some nouns, verbs, and adjectives."
> tagPOS(sentence, language = "en")
```

```
[1] "This/DT is/VBZ a/DT short/JJ sentence/NN consisting/VBG of/IN"
[2] "some/DT nouns,/JJ verbs,/NNS and/CC adjectives./VBG"
```

Sentence Detection

```
> s <- "This is a sentence. This another---but with dash-like
+       structures, and some commas. Maybe another with question
+       marks? Sure!"
> sentDetect(s, language = "en")
```

[1] http://opennlp.sourceforge.net/

```
[1] "This is a sentence. "
[2] "This another---but with dash-like\n      structures, and some commas. "
[3] "Maybe another with question\n      marks? "
[4] "Sure!"
```

Tokenizer

```
> s <- "¿Como se llama usted? El castellano es la lengua española oficial del Estado."
> tokenize(s, language = "es")
```

```
 [1] "¿"          "Como"       "se"         "llama"      "usted"
 [6] "?"          "El"         "castellano" "es"         "la"
[11] "lengua"     "española"   "oficial"    "del"        "Estado"
[16] "."
```

Enhancements to tm

The package provides transformations to enhance the **tm** package. The functions `tmTagPOS`, `tmSentDetect`, and `tmTokenize` are wrappers for above functions to be applied to plain text documents.