Introduction to the **openNLP** Package

Ingo Feinerer and Kurt Hornik

June 27, 2009

Abstract

The **openNLP** package.

Introduction

The **openNLP** package provides an R interface to openNLP (`http://opennlp.sourceforge.net/`).

Loading the Package

The package is loaded via

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

Models

To use **openNLP** for a certain language (currently, languages '`en`' (English), '`es`' (Spanish), '`de`' (German) and '`th`' (Thai) are supported), the corresponding openNLP model language packages (**openNLPmodels.en**, **openNLPmodels.es**, ...) need to be available. These packages "only" contain the corresponding model files obtained from `http://opennlp.sourceforge.net/models/`, and thus need to be installed, but not loaded.

At least the model language packages for English and Spanish, previously contained in a single **openNLPmodels** package, are available from CRAN.

Using the Package

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] "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.