NLPreadability: An R Package for Generating Readability Features

Abstract

This vignette shows how to use the **NLPreadability** package.

Keywords: readability, NLP, R.

1. Data

```
R> library("NLP")
R> library("tm")
```

To showcase the usage of this package we will use the **OneStopEnglish** corpus (Vajjala and Lučić 2018) and the **English Textbook** corpus (Islam 2015).

1.1. Installation

Both corpora are available and can be installed from `https://datacube.wu.ac.at/`.

```
R> dcube <- "https://datacube.wu.ac.at/"
R> install.packages("tm.corpus.enTextbook", repos = dcube, type = "source")
R> install.packages("tm.corpus.OneStopEnglish", repos = dcube, type = "source")
```

Both packages contain the corpus and the derived annotations and features. More information can be found in the corresponding `README` files.

2. Building the annotations

To build the annotations we use the **Stanford CoreNLP** (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky 2014) natural language software. **Stanford CoreNLP** is a Java software which can be accessed from within R through the packages **StanfordCoreNLP** (Hornik 2020) and **NLPclient** (Schwendinger and Hornik 2019).

2.1. Installation

StanfordCoreNLP

Package **Stanford CoreNLP** is available from the `https://datacube.wu.ac.at/` repository.

The main package can be installed with

```
R> install.packages("StanfordCoreNLP", repos = dcube, type = "source")
```

Additionally, pre-trained models for different languages can be installed.

```
R> pkgs <- available.packages(repos="https://datacube.wu.ac.at")
R> grep("StanfordCoreNLP", rownames(pkgs), value = TRUE)
```

```
[1] "StanfordCoreNLP"        "StanfordCoreNLPjars"    "StanfordCoreNLPjars.ar"
[4] "StanfordCoreNLPjars.de" "StanfordCoreNLPjars.es" "StanfordCoreNLPjars.fr"
[7] "StanfordCoreNLPjars.zh"
```

In order to install the English language models, one should use:

```
R> install.packages("StanfordCoreNLPjars", repos = dcube, type = "source")
```

NLPclient

Package **NLPclient** is available from CRAN.

```
R> install.packages("NLPclient")
```

More information about the installation of **NLPclient** can be found in the package README.

2.2. Annotation

To use the **NLPreadability** package the following annotators should be used:

```
R> annotators <- c("tokenize", "ssplit", "pos", "lemma", "ner",
+    "regexner", "truecase", "parse", "dcoref", "relation")
```

In the following we show the creation of the annotations for the **OneStopEnglish** corpus. Since **Stanford CoreNLP** needs considerable amounts of memory assigned to the virtual machine, we first increase the amount of memory **Java** is allowed to use.

```
R> # If you have more memory use more, my laptop has only 8GB.
R> options(java.parameters = "-Xmx6g", stringsAsFactors = FALSE )
```

We then load the **OneStopEnglish** corpus. The object `ose_corpus` contains a list of three corpora, one for each readability level.

```
R> library("StanfordCoreNLP")
R> library("tm.corpus.OneStopEnglish")
R> data("ose_corpus")
R> ose_corpus
```

```
$elementary
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 189

$intermediate
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 189

$advanced
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 189
```

The following command accesses the first text among the ones classified as advanced:

```
R> txt <- content(ose_corpus$advanced)[1]
```

In order to build the annotations the following code can be used:

```
R> p <- StanfordCoreNLP_Pipeline(annotators, control = list(nthreads = 1L))
R> annotate <- function(x) AnnotatedPlainTextDocument(x, p(x))

R> anno <- vector("list", sum(lengths(ose_corpus)))
R> k <- 1L
R> for (readability_level in names(ose_corpus)) {
+    corp <- ose_corpus[[readability_level]]
+    texts <- content(corp)
+    for (i in seq_along(texts)) {
+      names(anno)[k] <- sprintf("%s_%03i", substr(readability_level, 1, 3), i)
+      anno[[k]] <- annotate(texts[i])
+      k <- k + 1L
+    }
+  }
```

Since building the annotations is time consuming, pre-computed annotations can be loaded from the **tm.corpus.OneStopEnglish** (and the **tm.corpus.enTextbook** respectively) package.

```
R> data("ose_annotations")
```

The annotations for the first text in the advanced corpus can be accessed by:

```
R> ose_annotations["adv_001"]

$adv_001
<<AnnotatedPlainTextDocument>>
Metadata:  0
Annotations:  length: 745
Content:  chars: 3826
```

The names of the annotations consist of the first three letters of the readability level ("elementary", "intermediate" and "advanced") and the document id.

```
R> readability_level <- gsub("_.*", "", names(ose_annotations))
R> table(readability_level)

readability_level
adv ele int
189 189 189

R> id <- as.integer(gsub(".*_", "", names(ose_annotations)))
```

3. Building the features

The **NLPreadability** package simplifies the creation of features for readability prediction.

```
R> library(NLPreadability)
R> features <- lapply(ose_annotations, readability_features)
R> features <- do.call(rbind, features)
R> readability_level <- gsub("_.*", "", rownames(features))
R> y <- ordered(readability_level, levels = c("ele", "int", "adv"))
R> features <- cbind(readability = y, as.data.frame(as.matrix(features)))
```

For the **OneStopEnglish** corpus and the **English Textbook** corpus pre-computed features can be loaded from the corresponding packages.

```
R> library("tm.corpus.OneStopEnglish")
R> data("ose_features")
R> dim(ose_features)

[1] 567  96

R> library("tm.corpus.enTextbook")
R> data("entb_features")
R> dim(entb_features)

[1] 519  96
```

References

Hornik K (2020). *StanfordCoreNLP: Stanford CoreNLP Annotation*. R package version 0.1-5.

Islam MZ (2015). *Multilingual text classification using information-theoretic features*. Ph.D. thesis, Department of Computer Science. URL http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/year/2015/docId/38157.

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014). "The Stanford CoreNLP Natural Language Processing Toolkit." In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Schwendinger F, Hornik K (2019). *NLPclient: Stanford 'CoreNLP' Annotation Client*. R package version 1.0, URL https://CRAN.R-project.org/package=NLPclient.

Vajjala S, Lučić I (2018). "OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification." In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 297–304. Association for Computational Linguistics, New Orleans, Louisiana. doi:10.18653/v1/W18-0535. URL https://www.aclweb.org/anthology/W18-0535.

Affiliation:

Firstname Lastname
Affiliation
Address, Country
E-mail: name@address
URL: http://link/to/webpage/