

Introduction to the **RKEA** Package

Ingo Feinerer

Kurt Hornik

November 1, 2024

Abstract

A short introduction to the **RKEA** package.

Introduction

The **RKEA** package provides a R interface to Kea (<http://www.nzdl.org/Kea/>), a tool for keyword extraction in texts. See <https://code.google.com/p/kea-algorithm/> and <http://www.nzdl.org/Kea/Download/Kea-5.0-Readme.txt> for further information on Kea.

Note that Maui (<http://maui-indexer.googlecode.com/>), an algorithm for topic indexing, can be used for the same tasks as Kea, but offers additional features, including indexing using Wikipedia as a controlled vocabulary. See <https://www.airpair.com/nlp/keyword-extraction-tutorial> for a tutorial on NLP keyword extraction with Maui and RAKE (Rapid Automatic Keyword Extraction): note however that currently there is no R interface to Maui, nor an R implementation of RAKE.

Loading the Package

Before actually working we need to load the package:

```
> library("RKEA")
```

Creating a Keyword Extraction Model

Kea needs a keyword extraction model for keyword extraction. You can build your own models by manually indexing the keywords in a small set of texts, and then call `createModel()`.

```
> library("tm")
> data("crude")
> keywords <- list(c("Diamond", "crude oil", "price"),
+                 c("OPEC", "oil", "price"),
+                 c("Texaco", "oil", "price", "decrease"),
+                 c("Marathon Petroleum", "crude", "decrease"),
+                 c("Houston Oil", "revenues", "decrease"),
+                 c("Kuwait", "OPEC", "quota"))
> tmpdir <- tempfile()
> dir.create(tmpdir)
> model <- file.path(tmpdir, "crudeModel")
> createModel(crude[1:6], keywords, model)
```

Please note that we just wrap the functionality of the original Kea program which always uses files for in- and output (and that is the reason you also need to use a directory in R as shown in the above example). We deliberately decided not to modify the Kea Java archive shipped with this R package for compatibility reasons. However this may induce some warnings in R (e.g., because some internal Kea paths might not be available) but nevertheless you should get the full functionality out of it.

Keyword Extraction

Once you have a Kea model you can extract keywords from texts.

```
> extractKeywords(crude, model)

[[1]]
 [1] "Diamond"          "cut"              "cut its contract"
 [4] "reduction"        "contract"         "crude oil"
 [7] "dlrs a barrel"    "prices"           "today"
[10] "crude"

[[2]]
 [1] "OPEC"          "problems"  "production" "analysts"  "Energy"
 [6] "meet"         "OPEC's"   "oil prices" "meeting"   "June"

[[3]]
 [1] "Texaco Canada"          "Texaco"
 [3] "Canada"                "it lowered the contract price"
 [5] "crude oil"             "crude"
 [7] "price"                 "it"
 [9] "oil"                   "Edmonton/Swann"

[[4]]
 [1] "it reduced the contract price" "reduced"
 [3] "grades of crude"             "grades"
 [5] "crude"                       "price"
 [7] "it"                          "West Texas Intermediate"
 [9] "posted"                      "posted price"

[[5]]
 [1] "future net"          "future"          "reserves"
 [4] "estimates"          "Trust said"     "Trust"
 [7] "study"              "future net revenues" "net"
[10] "revenues"

[[6]]
 [1] "Kuwait"          "OPEC"          "bpd"          "Minister"
 [5] "Sheikh"          "Sheikh Ali"   "Ali"          "members"
 [9] "international"  "quota"

[[7]]
 [1] "says"          "Indonesia"  "economy"  "government"  "report says"
```

[6] "report" "appears" "measures" "sector" "investment"

[[8]]
 [1] "higher levels" "riyal" "deposits" "higher"
 [5] "OPEC" "market" "yesterday's" "said"
 [9] "yesterday" "quotes"

[[9]]
 [1] "billion" "budget" "billion riyals"
 [4] "riyals" "government" "Sheikh Abdul-Aziz"
 [7] "Abdul-Aziz" "expenditure" "decline"
 [10] "Sheikh"

[[10]]
 [1] "Saudi" "accord" "Nazer" "commitment" "OPEC"
 [6] "OPEC accord" "SPA" "free" "free market" "market"

[[11]]
 [1] "Saudi" "exports" "January"
 [4] "bpd" "output" "fell"
 [7] "average" "Ju'aymah terminals" "Ju'aymah"
 [10] "February"

[[12]]
 [1] "official" "oil ministers" "ministers" "Gulf"
 [5] "Arab" "states" "Emirates" "crude oil"
 [9] "crude" "oil"

[[13]]
 [1] "Saudi" "commitment" "OPEC accord"
 [4] "OPEC" "Nazer" "commitment to last"
 [7] "OPEC accord to boost" "accord" "oil prices"
 [10] "kingdom's"

[[14]]
 [1] "oil minister said" "OPEC"
 [3] "oil minister" "meeting"
 [5] "oil prices" "prices"
 [7] "oil" "pumping above its OPEC"
 [9] "daily" "pumping"

[[15]]
 [1] "power" "port" "closed"
 [4] "lines" "oil" "ship"
 [7] "nuclear power plant" "nuclear" "nuclear power"
 [10] "plant"

[[16]]
 [1] "group" "strategic"
 [3] "mln barrels" "oil prices on the domestic"

```

[5] "reserve"                "present"
[7] "U.S"                    "mln"
[9] "barrels"                "prices"

[[17]]
[1] "group"                  "strategic"
[3] "mln barrels"           "oil prices on the domestic"
[5] "reserve"                "U.S"
[7] "study"                  "present"
[9] "policy"                 "industry"

[[18]]
[1] "Union"                  "Union Oil Co"      "Union Oil"
[4] "Union Oil Co said" "posted"             "posted prices"
[7] "lowered"            "Corp's"             "crude oil"
[10] "dlrs"

[[19]]
[1] "NYMEX"                  "Exchange"
[3] "futures"                "transaction"
[5] "change"                 "hold a futures"
[7] "hold a futures position" "futures position"
[9] "position"               "traders"

[[20]]
[1] "January"
[2] "mln barrels"
[3] "pct"
[4] "mln"
[5] "Yacimientos Petroliferos"
[6] "Yacimientos Petroliferos Fiscales"
[7] "Yacimientos"
[8] "Petroliferos"
[9] "Petroliferos Fiscales"
[10] "Fiscales"

> unlink(tmpdir, recursive = TRUE)

```

Working with Controlled Vocabularies

The data used for the keyword extraction tutorial with Maui and RAKE can be downloaded from <https://maui-indexer.googlecode.com/files/fao780.tar.gz>; the AGROVOC Agricultural Thesaurus can be obtained from <http://www.nzdl.org/Kea/Download/vocabularies/agrovoc.skos.zip> (SKOS format) or <http://www.nzdl.org/Kea/Download/vocabularies/agrovoc.text.zip> (text format).

With the data unpacked to subdirectory `fao780` and `agrovoc.skos.zip` unzipped in the working directory, one can use

```

> txts <- Sys.glob(file.path("fao780", "*.txt"))
> keys <- sub("txt$", "key", txts)

```

```
> txts <- lapply(txts, readLines)
> keys <- lapply(keys, readLines)
> build <- seq_len(100)
> xtrct <- seq(101, 105)
> model <- "fao780_model"
> createModel(txts[build], keys[build], model, "agrovoc", "skos")
> extractKeywords(txts[xtrct], model, "agrovoc", "skos")
```

to build a keyword model using the first 100 texts, and use the model to extract the keywords from the next 5 texts.