Text Plots

Jan Wijffels

Abstract

The textplot R package allows one to visualise complex relations in texts. This is done by providing functionalities for displaying text co-occurrence networks, text correlation networks, dependency relationships as well as text clustering. In this vignette, some example visualisations of these are shown.

Keywords: Text, network, co-occurrence, correlation, text clustering, dependency parsing, visualisation.

1. General

1.1. Overview

The package allows you to visualise

- Text frequencies
- Text correlations
- Text cooccurrences
- Text clusters
- Dependency parsing results

Source code repository

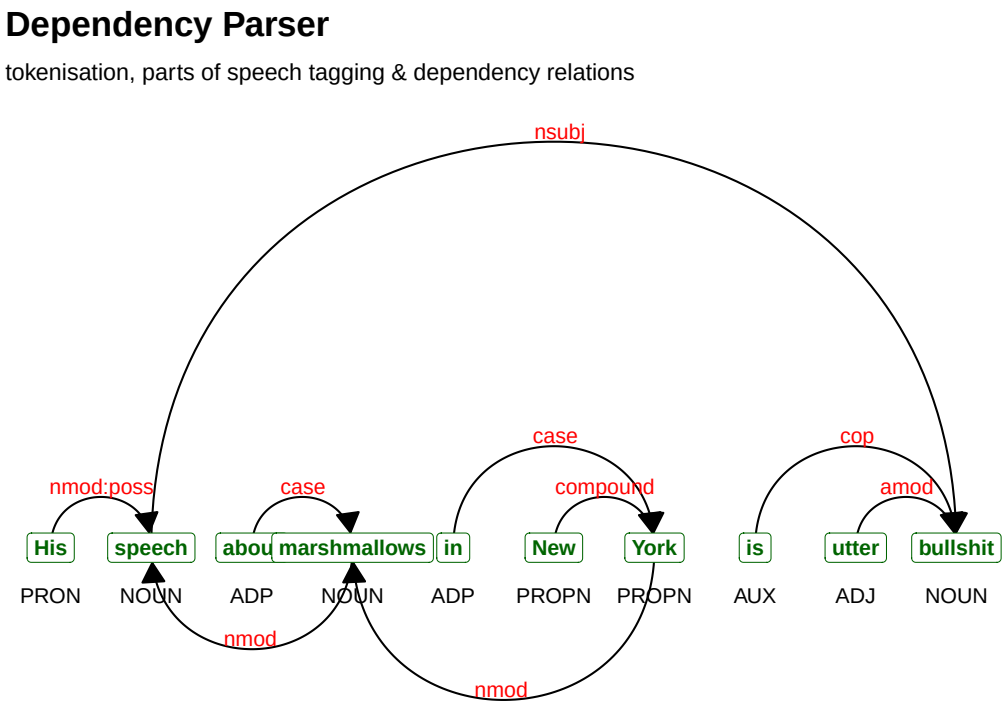
The source code of the package is on github at `https://github.com/bnosac/textplot`.
The R package is distributed under the GPL-2 license.

2. Example visualisations

2.1. Dependency Parser

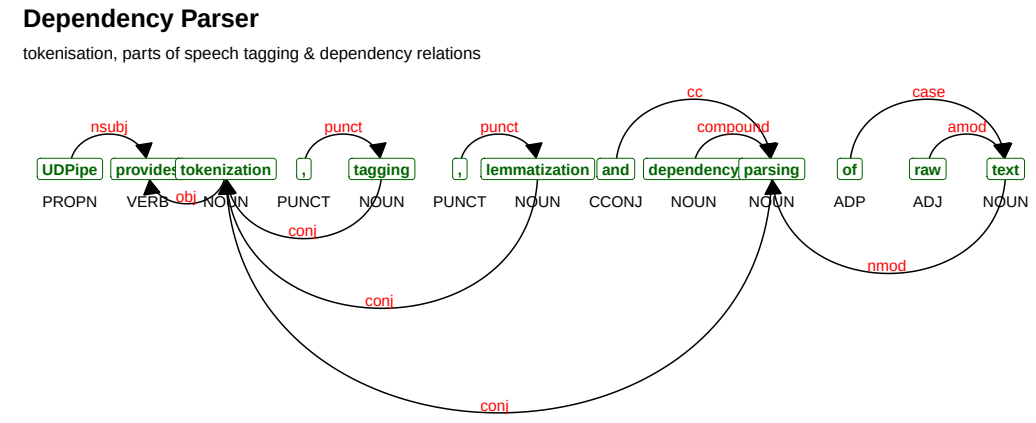
Example 1

```
library(udpipe)
library(textplot)
library(ggraph)
library(igraph)
x <- udpipe("His speech about marshmallows in New York is utter bullshit",
            "english")
plt <- textplot_dependencyparser(x, size = 4)
plt
```



Example 2

```
x <- udpipe("UDPipe provides tokenization, tagging, lemmatization and
             dependency parsing of raw text", "english")
plt <- textplot_dependencyparser(x, size = 4)
plt
```

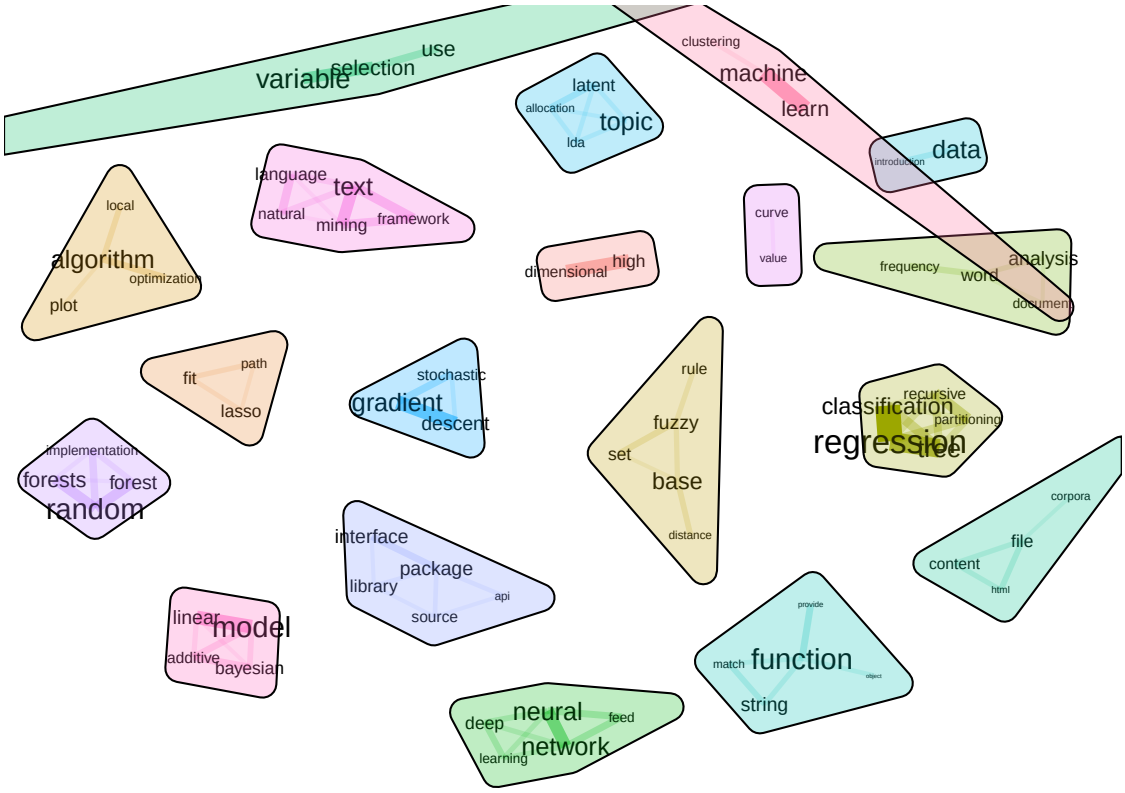


2.2. Biterm Topic Model plots

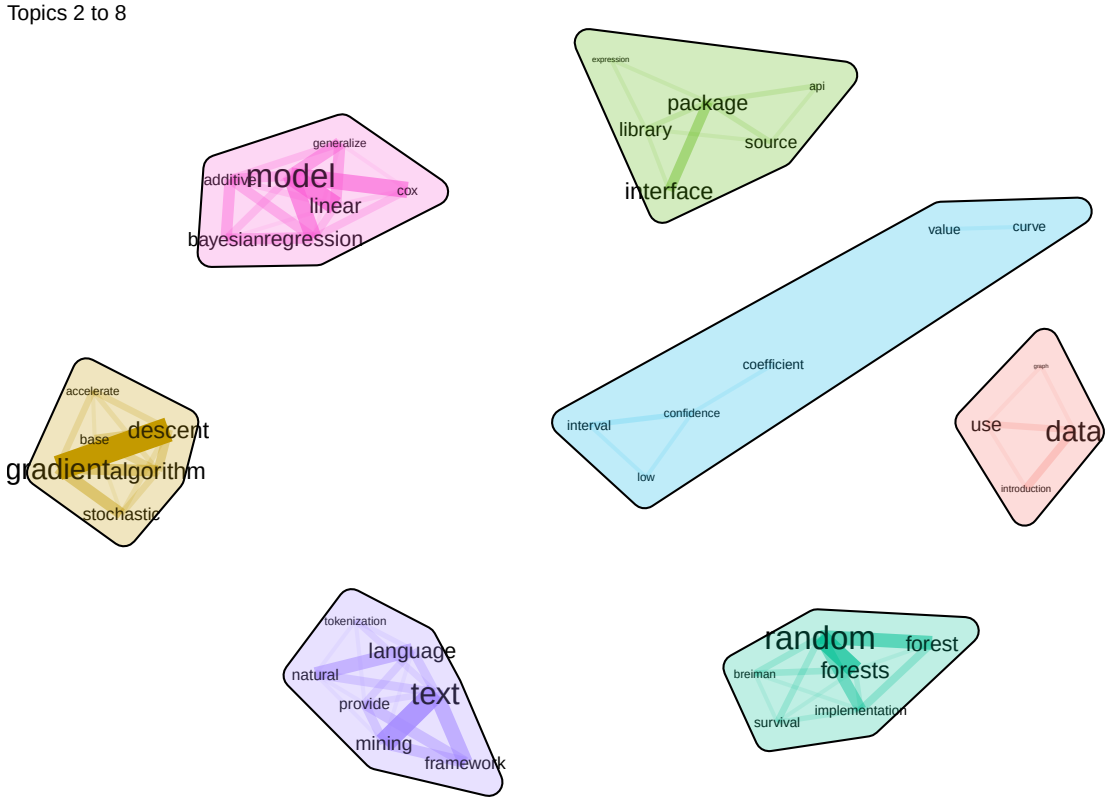
Example 1

```
library(BTM)
library(ggraph)
library(concaveman)
library(igraph)
data(example_btm, package = 'textplot')
model <- example_btm
plt <- plot(model, title = "BTM model", top_n = 5)
plt
```

BTM model




```
plt <- plot(model, title = "Biterm topic model", subtitle = "Topics 2 to 8",
            which = 2:8, top_n = 7)
plt
```

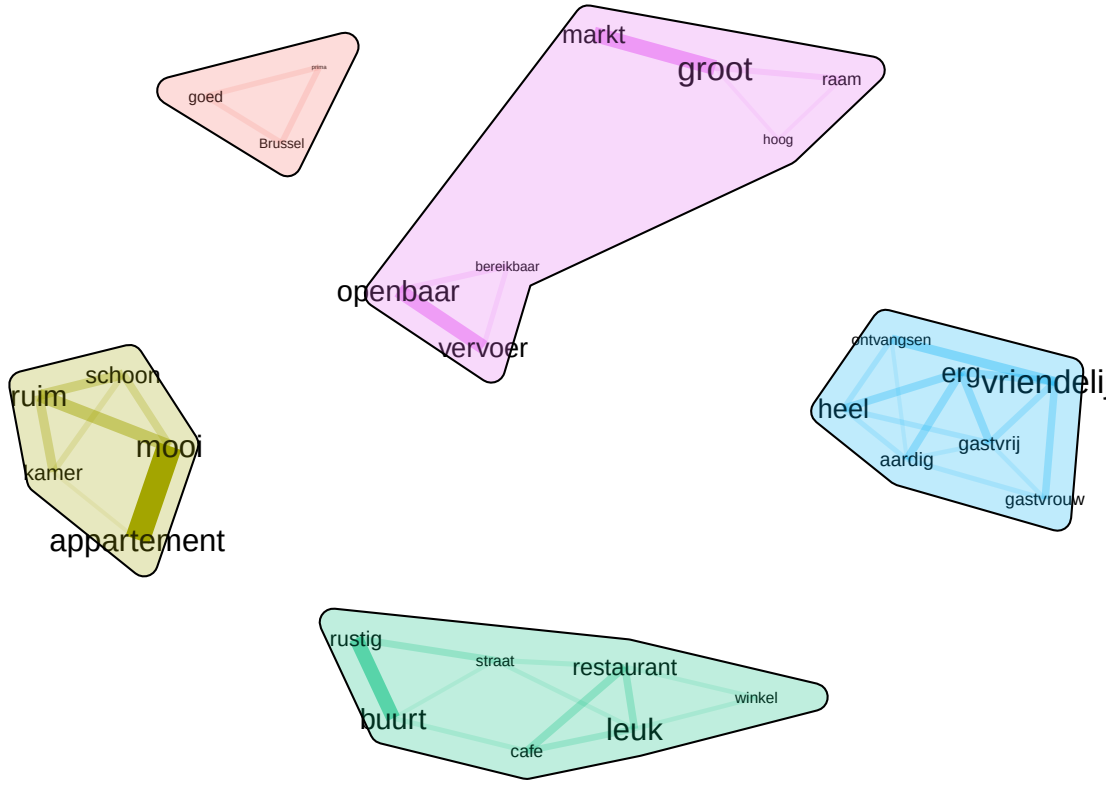


Example 2

```
library(BTM)
library(data.table)
library(udpipe)
library(igraph)
## Annotate text with parts of speech tags
data("brussels_reviews", package = "udpipe")
anno <- subset(brussels_reviews, language %in% "nl")
anno <- data.frame(doc_id = anno$id, text = anno$feedback, stringsAsFactors = FALSE)
anno <- udpipe(anno, "dutch", trace = 10)
## Get cooccurrences of nouns / adjectives and proper nouns
biterms <- as.data.table(anno)
biterms <- biterms[, cooccurrence(x = lemma,
                                  relevant = upos %in% c("NOUN", "PROPN", "ADJ"),
                                  skipgram = 2),
                   by = list(doc_id)]
## Build the BTM model
set.seed(123456)
x <- subset(anno, upos %in% c("NOUN", "PROPN", "ADJ"))
x <- x[, c("doc_id", "lemma")]
model <- BTM(x, k = 5, beta = 0.01, iter = 2000, background = TRUE,
```

```
                biterms = biterms, trace = 100)
plt <- plot(model)
plt
```

Biterm topic model



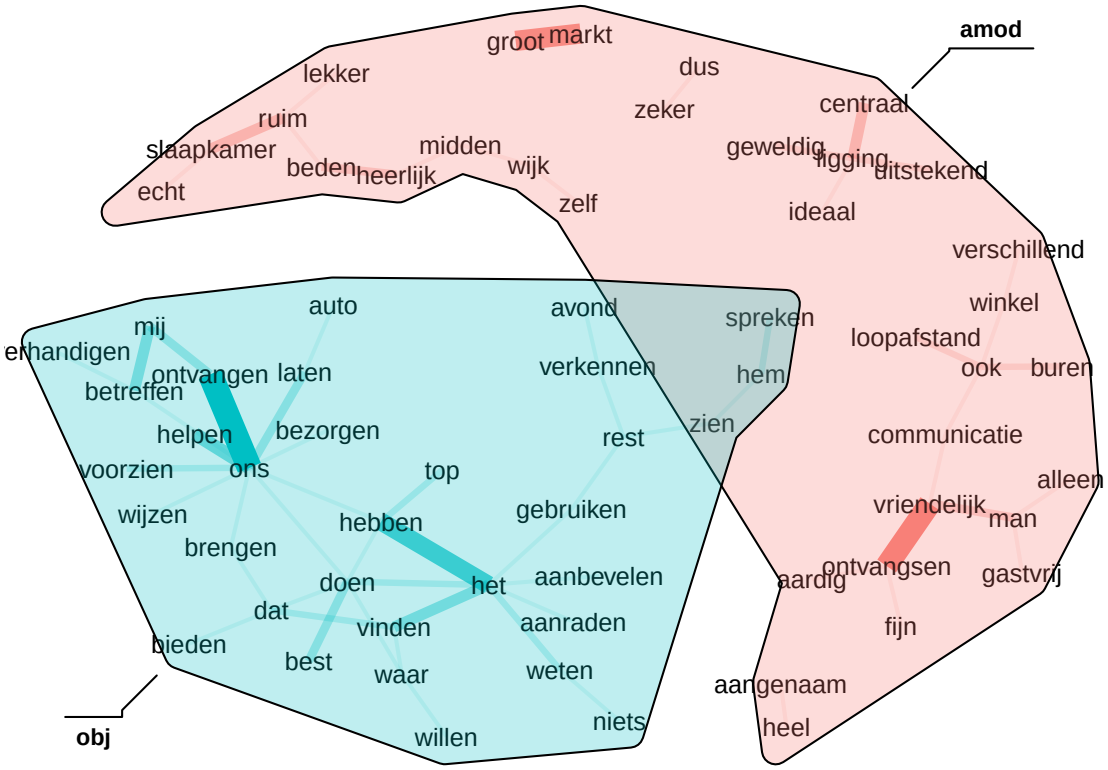
2.3. Biterm relationships

Example showing objects of verbs and adjectives modifying nouns

```
library(udpipe)
library(data.table)
library(ggraph)
x <- merge(anno, anno,
           by.x = c("doc_id", "paragraph_id", "sentence_id", "head_token_id"),
           by.y = c("doc_id", "paragraph_id", "sentence_id", "token_id"),
           all.x = TRUE, all.y = FALSE, suffixes = c("", "_parent"), sort = FALSE)
x <- subset(x, dep_rel %in% c("obj", "amod"))
x$topic <- factor(x$dep_rel)
topiclabels <- levels(x$topic)
x$topic <- as.integer(x$topic)
## Construct biterms/terminology inputs to the plot
biterms <- data.frame(term1 = x$lemma, term2 = x$lemma_parent,
                      topic = x$topic, stringsAsFactors = FALSE)
terminology <- document_term_frequencies(x, document = "topic",
                                         term = c("lemma", "lemma_parent"))
terminology <- document_term_frequencies_statistics(terminology)
terminology <- terminology[order(terminology$tf_idf, decreasing = TRUE), ]
terminology <- terminology[, head(.SD, 50), by = list(topic = doc_id)]
terminology <- data.frame(topic = terminology$topic,
                          token = terminology$term,
                          probability = 1, stringsAsFactors = FALSE)
plt <- textplot_bitermclusters(terminology, biterms,
                               labels = topiclabels,
                               title = "Objects of verbs and adjectives modifying nouns
                               subtitle = "Top 50 by group")
plt
```

Objects of verbs and adjectives modifying nouns

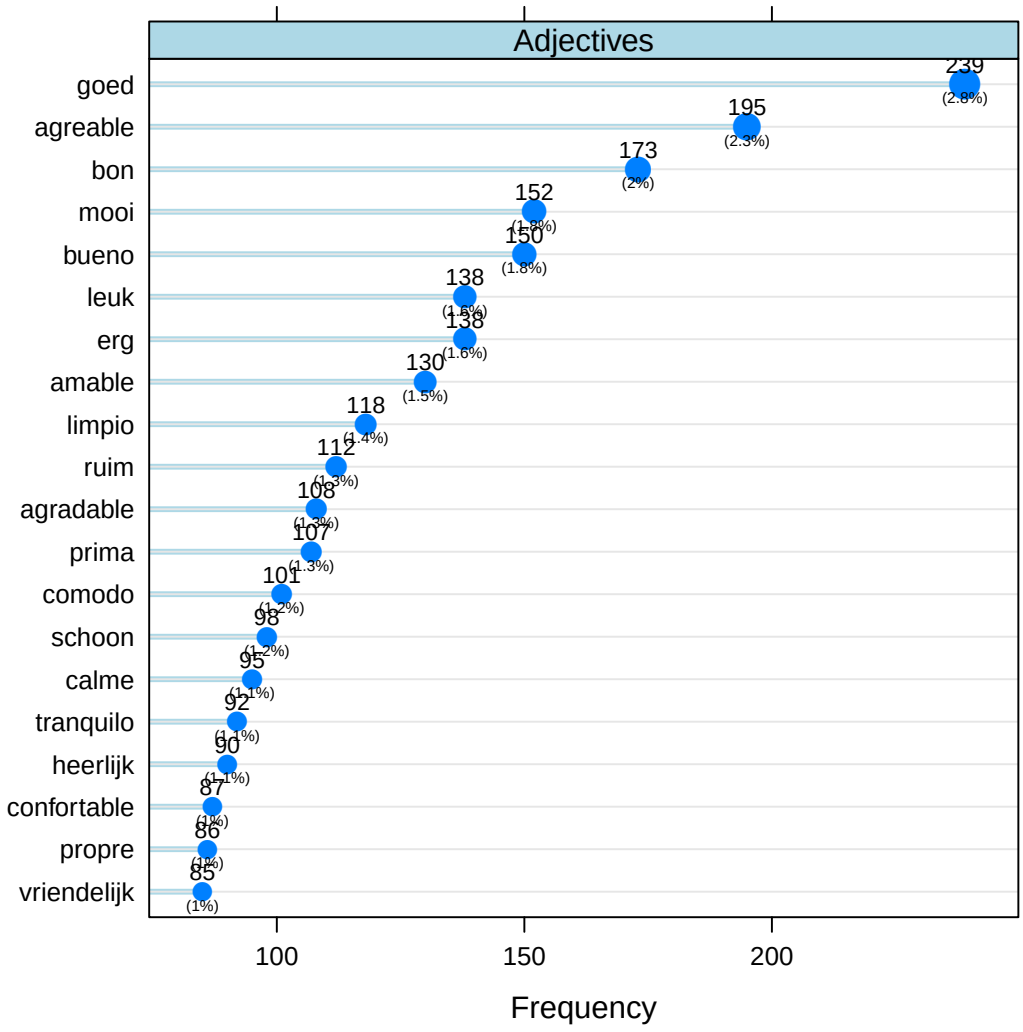
Top 50 by group



2.4. Bar plots

Example showing frequency of adjectives

```
library(udpipe)
data("brussels_reviews_anno", package = "udpipe")
x   <- subset(brussels_reviews_anno, xpos %in% "JJ")
x   <- sort(table(x$lemma))
plt <- textplot_bar(x, top = 20,
                    panel = "Adjectives", xlab = "Frequency",
                    col.panel = "lightblue", cextext = 0.75,
                    addpct = TRUE, cexpct = 0.5)
plt
```



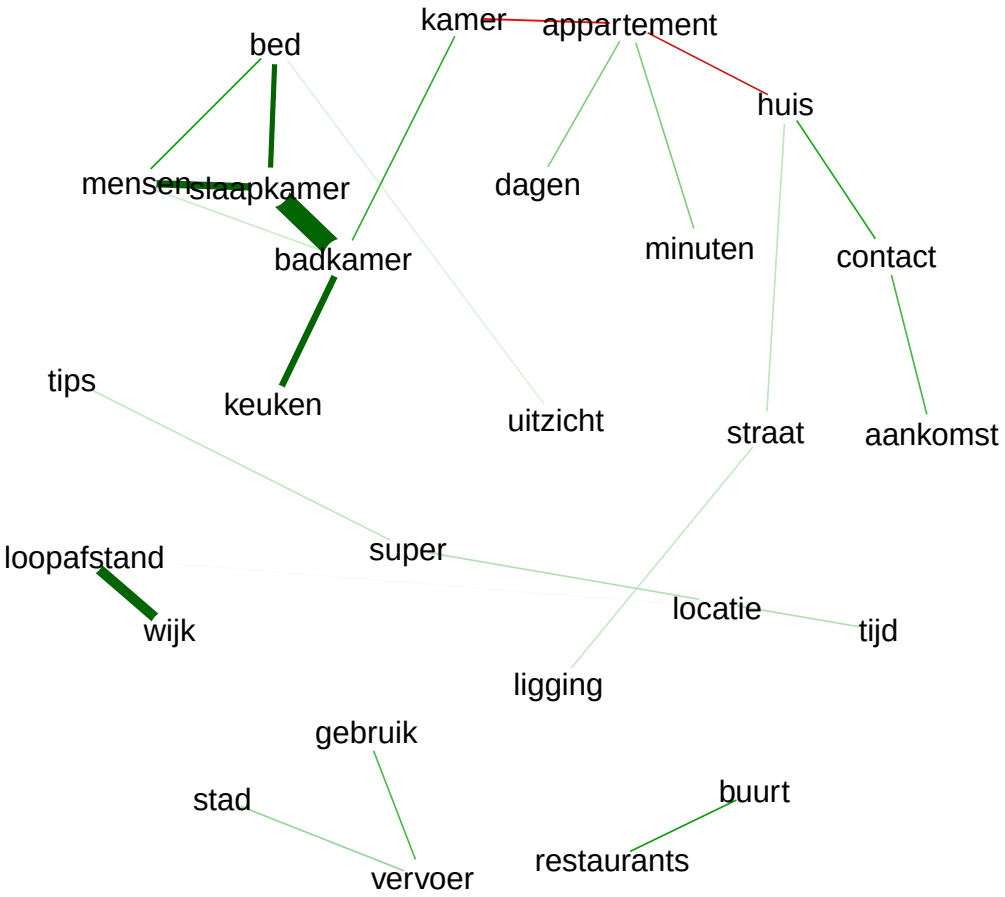
2.5. Correlation of texts

Top correlations above a certain threshold

```
library(graph)
library(Rgraphviz)
library(udpipe)
dtm <- subset(anno, upos %in% "ADJ")
dtm <- document_term_frequencies(dtm, document = "doc_id", term = "lemma")
dtm <- document_term_matrix(dtm)
dtm <- dtm_remove_lowfreq(dtm, minfreq = 5)
textplot_correlation_lines(dtm, top_n = 25, threshold = 0.01, lwd = 5, label = TRUE)
```

Correlations which are non-zero after fitting a glasso model

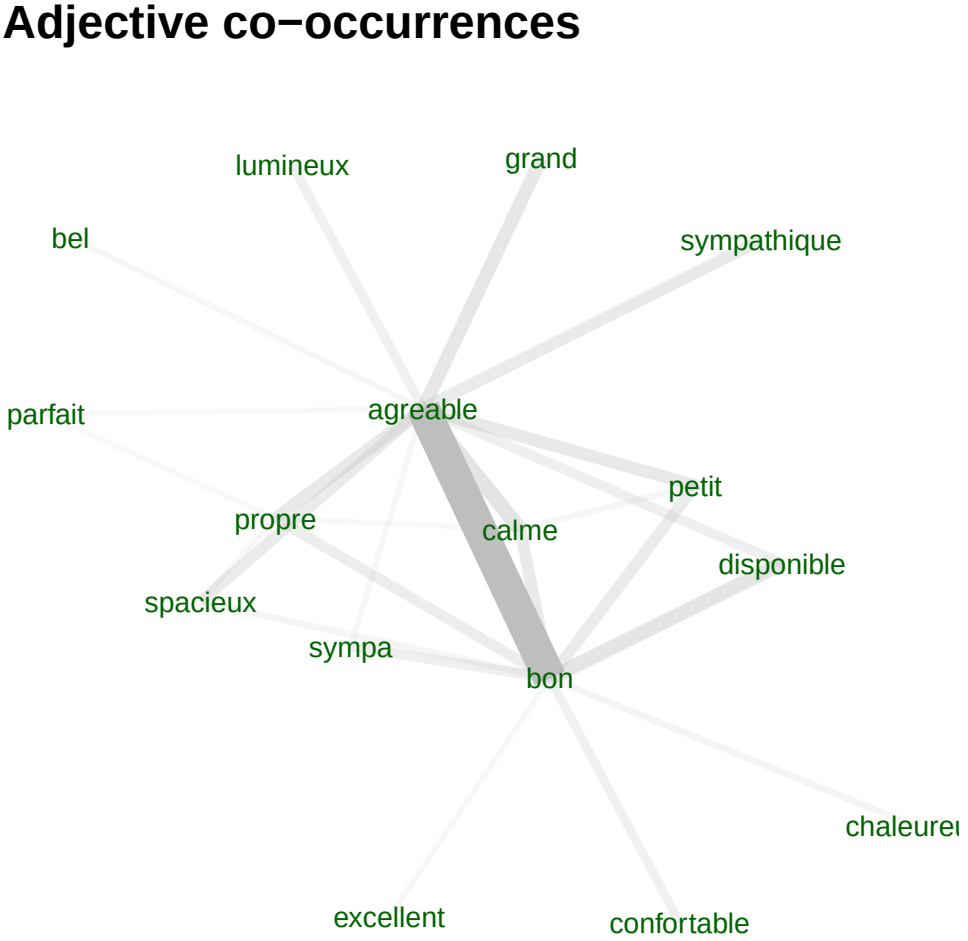
```
library(glasso)
library(qgraph)
library(udpipe)
dtm <- subset(anno, upos %in% "NOUN")
dtm <- document_term_frequencies(dtm, document = "doc_id", term = "token")
dtm <- document_term_matrix(dtm)
dtm <- dtm_remove_lowfreq(dtm, minfreq = 20)
dtm <- dtm_remove_tfidf(dtm, top = 100)
term_correlations <- dtm_cor(dtm)
textplot_correlation_glasso(term_correlations, exclude_zero = TRUE)
```



2.6. Co-occurrence of texts

Example showing adjectives occurring in the same document

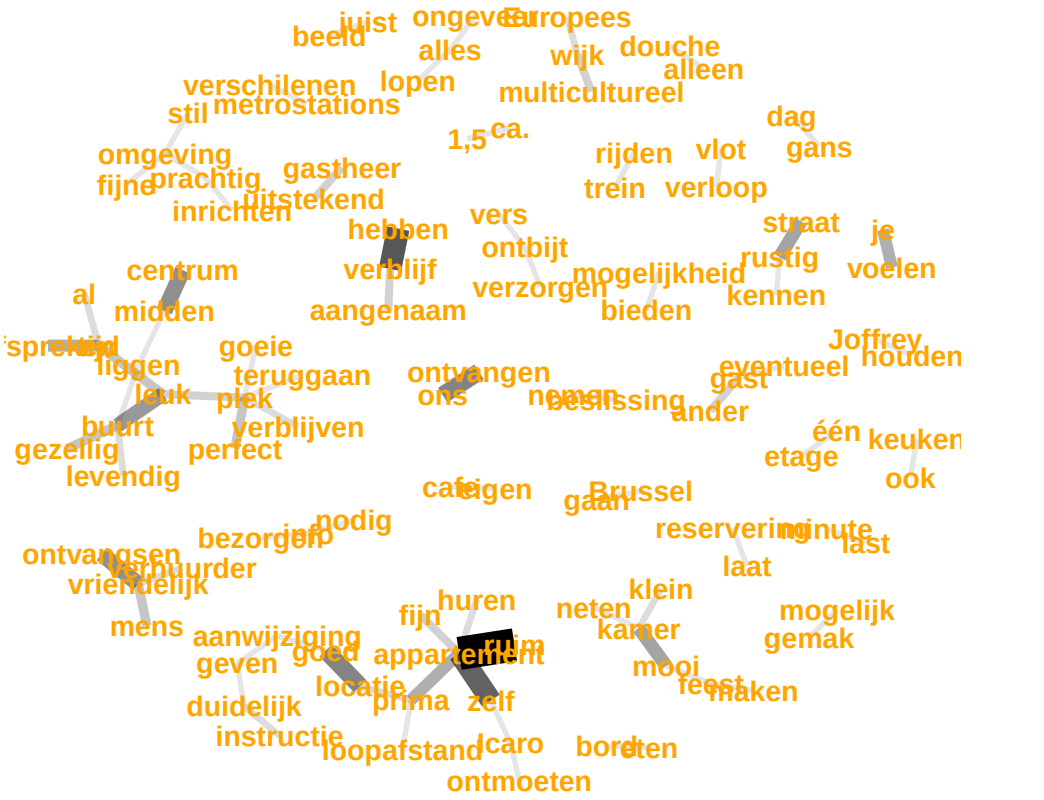
```
library(udpipe)
library(ggraph)
library(igraph)
data(brussels_reviews_anno, package = 'udpipe')
x <- subset(brussels_reviews_anno, xpos %in% "JJ" & language %in% "fr")
x <- cooccurrence(x, group = "doc_id", term = "lemma")
plt <- textplot_cooccurrence(x,
                             title = "Adjective co-occurrences", top_n = 25)
plt
```



Example showing objects of verbs / adjectives modifying nouns on our annotated dataset

```
library(udpipe)
library(ggraph)
library(igraph)
library(data.table)
data("brussels_reviews", package = "udpipe")
anno <- subset(brussels_reviews, language %in% "nl")
anno <- data.frame(doc_id = anno$id, text = anno$feedback, stringsAsFactors = FALSE)
anno <- udpipe(anno, "dutch", trace = 10)
```

```
biterms <- merge(anno, anno,
              by.x = c("doc_id", "paragraph_id", "sentence_id", "head_token_id"),
              by.y = c("doc_id", "paragraph_id", "sentence_id", "token_id"),
              all.x = TRUE, all.y = FALSE, suffixes = c("", "_parent"), sort = FALSE)
biterms <- setDT(biterms)
biterms <- subset(biterms, dep_rel %in% c("obj", "amod"))
biterms <- biterms[, list(cooc = .N), by = list(term1 = lemma, term2 = lemma_parent)]
plt <- textplot_cooccurrence(biterms,
                             title = "Objects of verbs and Adjectives modifying nouns",
                             vertex_color = "orange", edge_color = "black",
                             fontface = "bold")
plt
```



Affiliation:

BNOSAC - Open Analytical Helpers
E-mail: jwijffels@bnosac.be
URL: http://www.bnosac.be