
IDB
Inter-American
Development Bank

CRM Opportunity Object

Descriptive and Text mining Analysis

Romina Pardo

Text mining analysis: Procedure and Overview

I. Load Libraries and data sets.

II. Filter out all records with missing description.

III. View frequencies of languages in description field.

```
##
##        catalan         danish        english         french        frisian
##              6              1             80              4              1
##          latin middle_frisian         nepali      rumantsch        spanish
##              1              1              1              2            121
```

IV. Create separate data sets for descriptions in english and spanish. Prepare description Field for Text Mining Analysis: lowercase, remove stopwords.

V. Download Description in languages other than english, spanish or french. (./OppDesc.xlsx)

Previous results of Descriptive Analysis

1. 75% of Opportunities have been generated by Ecuador, Jamaica, Mexico and Uruguay.
2. IDB has initiated 50%. 25% has NA initiator.
3. 26% of Opportunities are in Education Sector, 20% are NA. The 3rd postion is shared by multisector, Financial Markets and Water and Sanitation, each with 10% of Opportunities.
4. 29% of Opportunities are Closed (10% Lost and 9% won).

   Download Descriptive Analysis. (./Opportunity.xlsx)

English POS Tagging

Use of pre-trained open- sourced models provided by UDpipe Community:

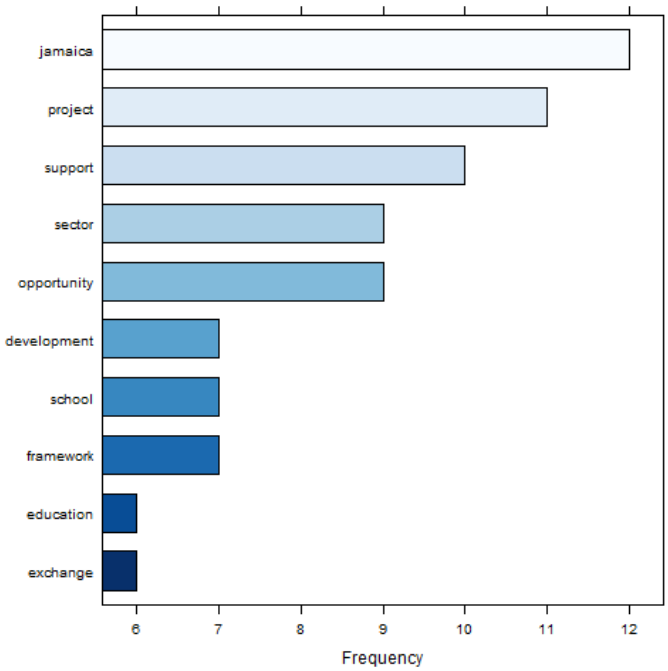
https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html (https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html)

UPOS (Universal Parts of Speech) frequency of occurrence:

```
##    key freq freq_pct
## 1 NOUN  780 50.32258
## 2 VERB  227 14.64516
## 3  ADJ  194 12.51613
```

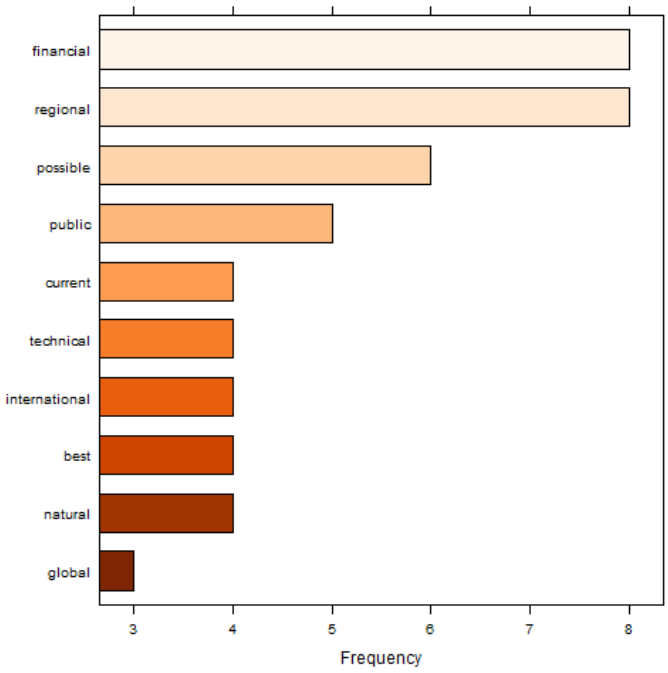
English POS: Nouns

Most Occurring Nouns in Descriptions



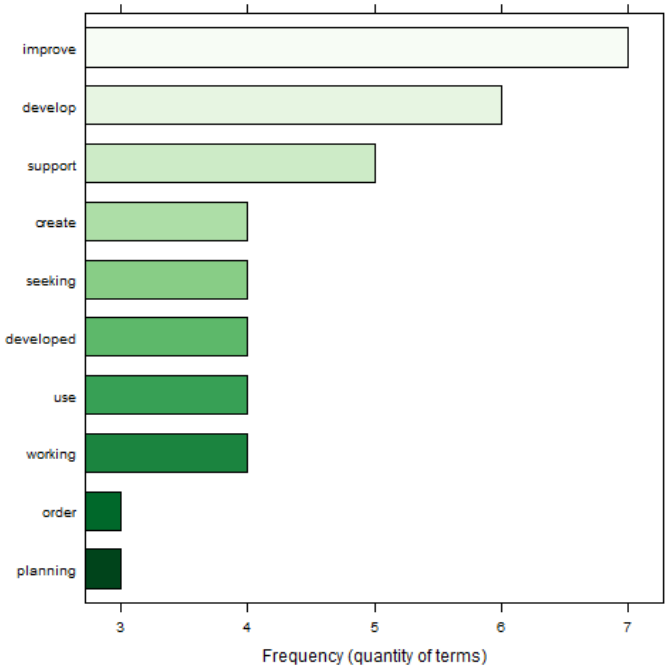
English POS: Adjectives

Most Occurring Adjectives in Descriptions



English POS: Verbs

Most Occurring Verbs in Descriptions



RAKE: Rapid Automatic Keyword Extraction algorithm

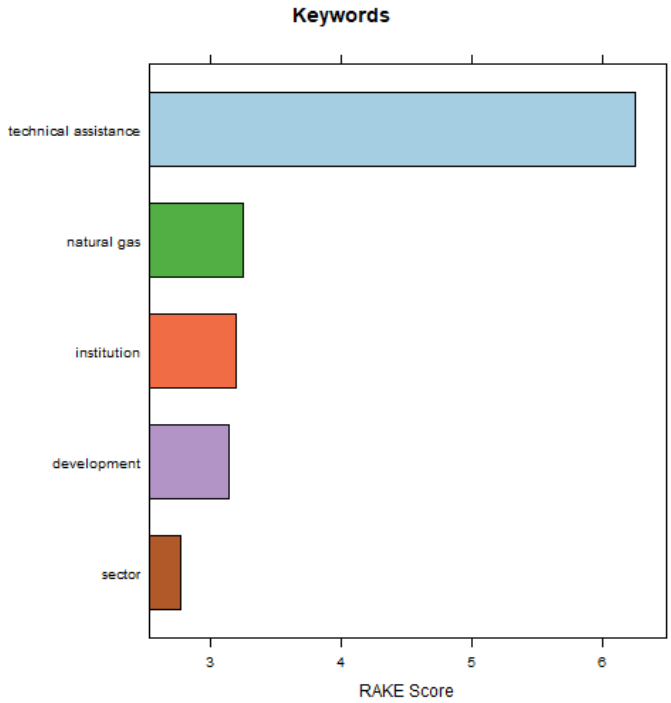
- Unsupervised algorithm that scores key phrases in a body of text by analyzing the frequency of each word appearance and its co-occurrence with other words in the text.

- It looks for a contiguous sequence of relevant words searching for keywords.

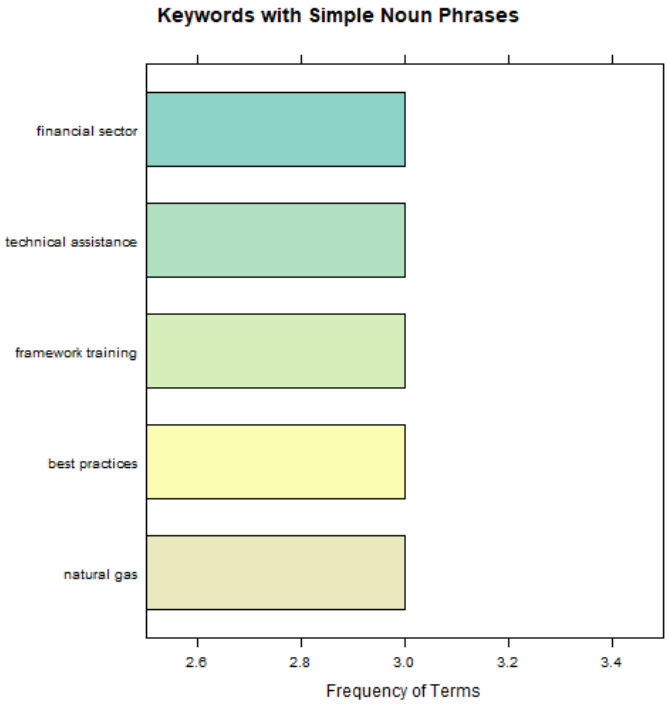
- For each word of any candidate keyword,it calculates a score which is the ratio of the word degree (how many times it co-occurs with other words) to the word frequency.

- A RAKE score for the full candidate keyword is calculated by summing up the scores of each of the words which define the candidate keyword

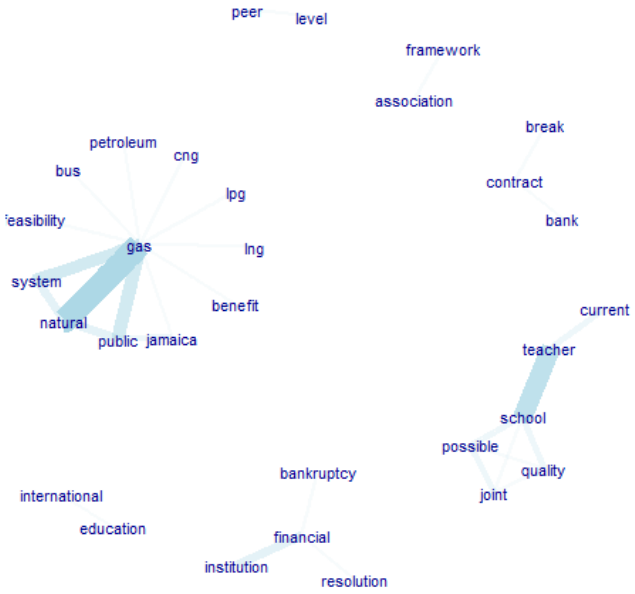
English keywords Using RAKE



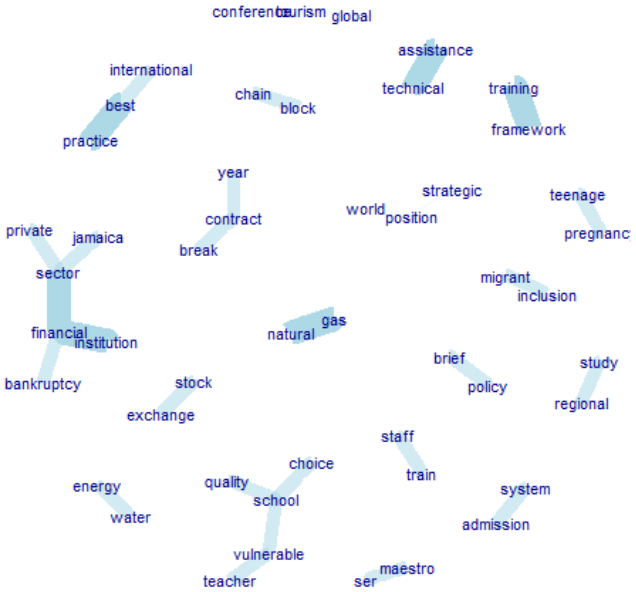
English top noun - verbs pairs as keyword pairs



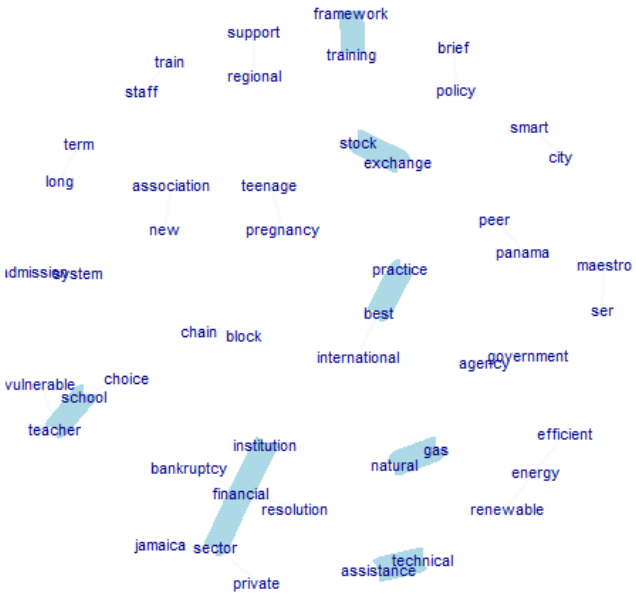
Co-occurrences: Frequency of words -nouns & adjectives - in the same sentence



Co-occurrences: Frequency of words -nouns & adjectives - following one another

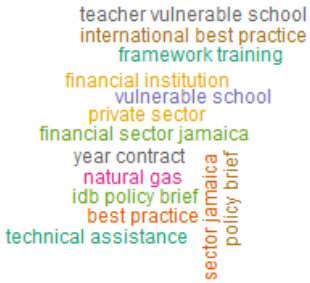


Co-occurrences: Frequency of words -nouns & adjectives - following one another skipping up to 2 words in between



English Textrank (word network ordered by Google Pagerank)

Unigram and ngrams with frequency over 2



Spanish POS Tagging

Use of pre-trained open- sourced models provided by UDpipe Community:

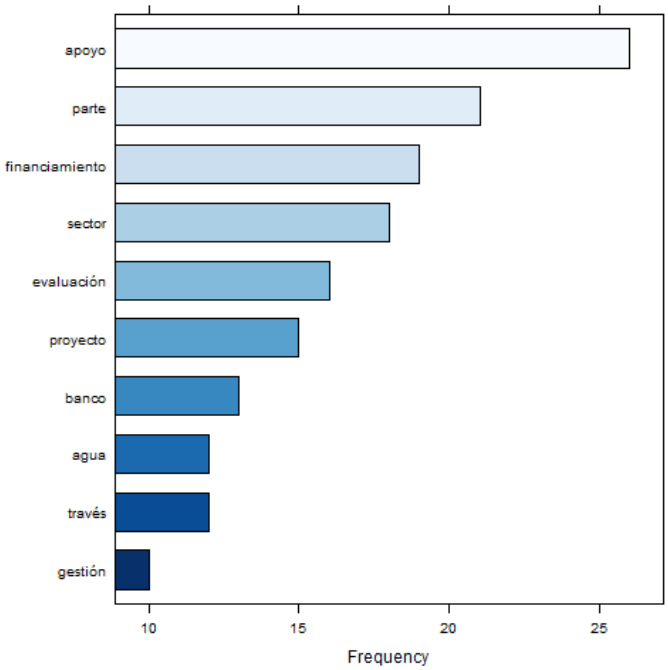
https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html (https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html)

UPOS (Universal Parts of Speech) frequency of occurrence:

```
##    key freq freq_pct
## 1 NOUN 1420 37.24102
## 2  ADJ  743 19.48597
## 3 VERB  586 15.36848
```

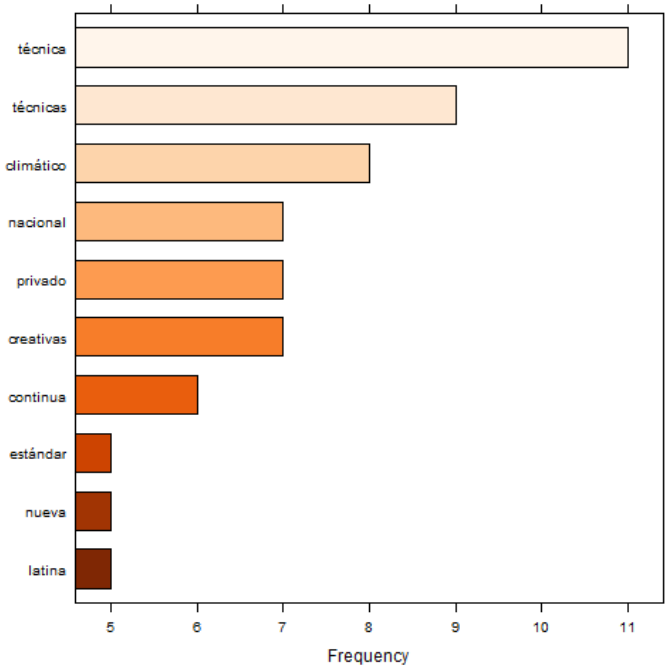
Spanish POS: Nouns

Most Occurring Nouns in Descriptions



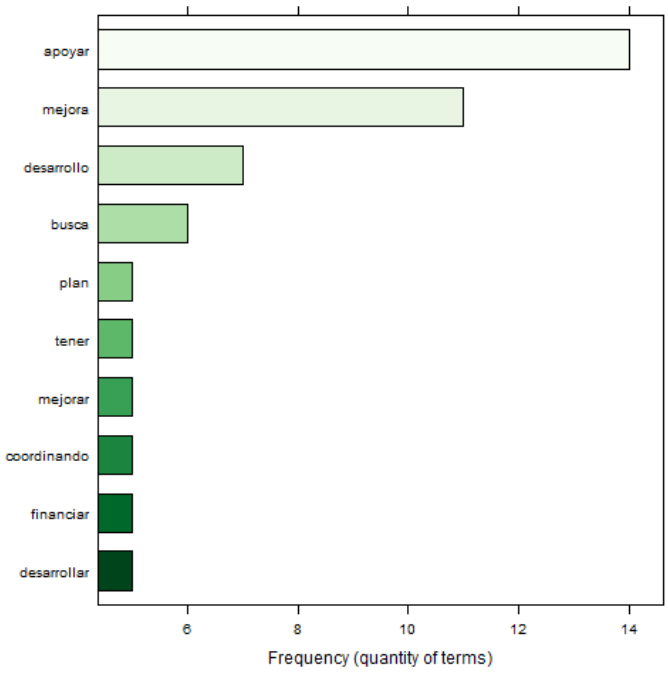
Spanish POS: Adjectives

Most Occurring Adjectives in Descriptions

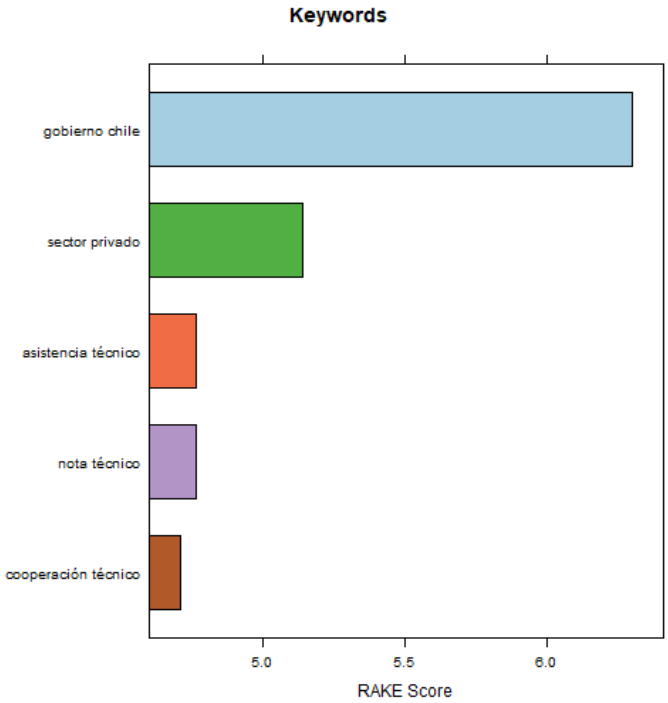


Spanish POS: Verbs

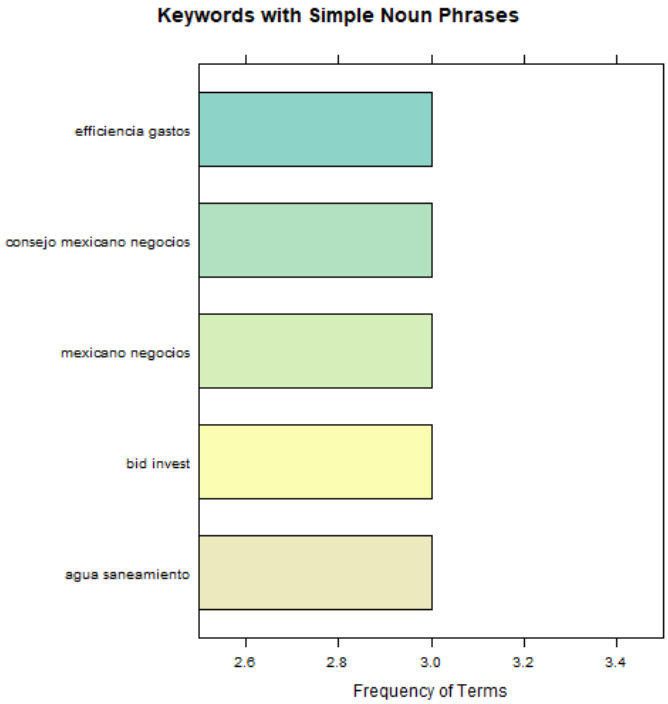
Most Occurring Verbs in Descriptions



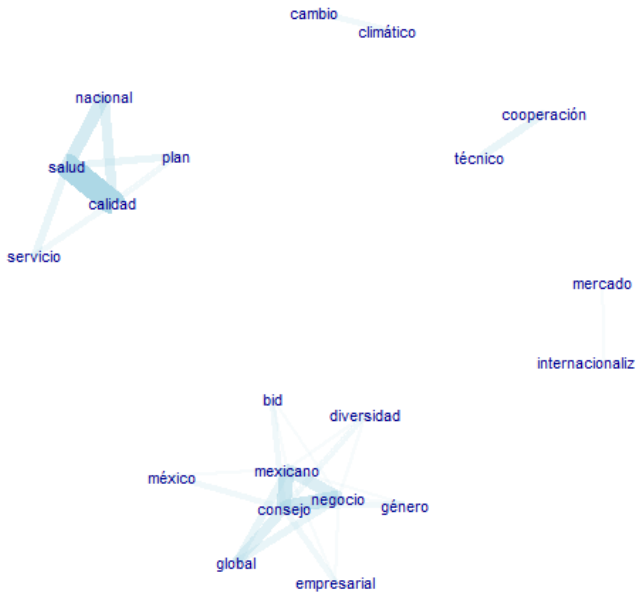
Spanish keywords Using RAKE



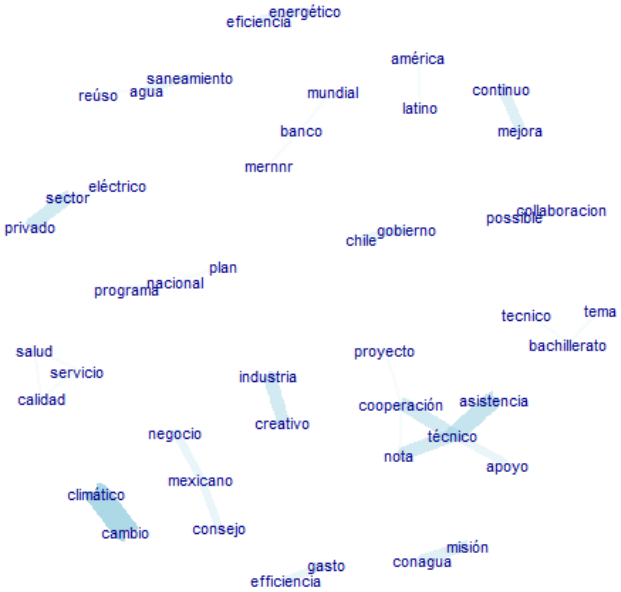
Spanish top noun - verbs pairs as keyword pairs



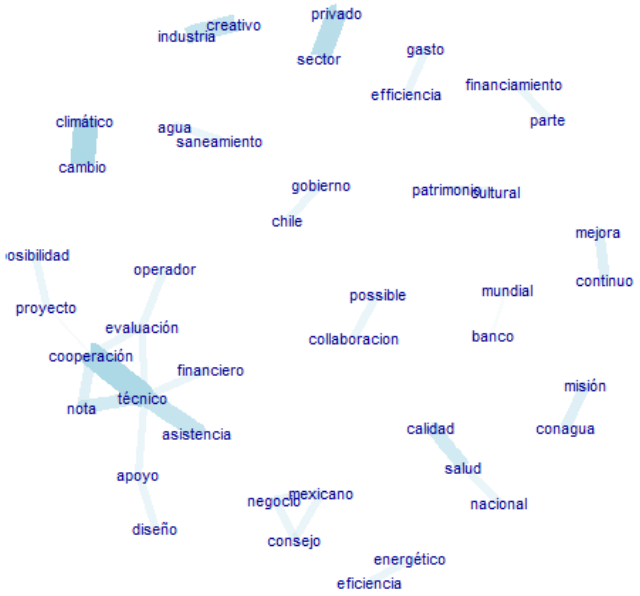
Co-occurrences: Frequency of words -nouns & adjectives - in the same sentence



Co-occurrences: Frequency of words -nouns & adjectives - following one another



Co-occurrences: Frequency of words -nouns & adjectives - following one another skipping up to 2 words in between



Spanish Textrank (word network ordered by Google Pagerank)

Unigram and ngrams with frequency over 2