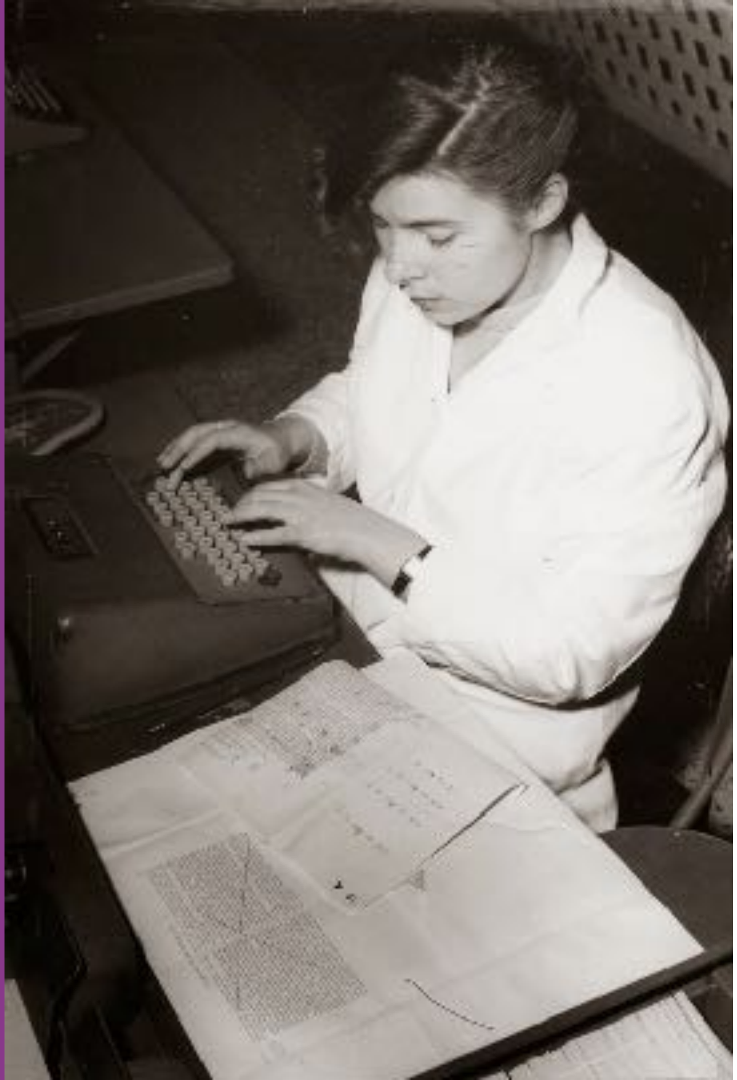


```
library(dplyr)

rladies_global %>%
  filter(city == 'Santiago')
```

Análisis de textos con R

Riva Quiroga

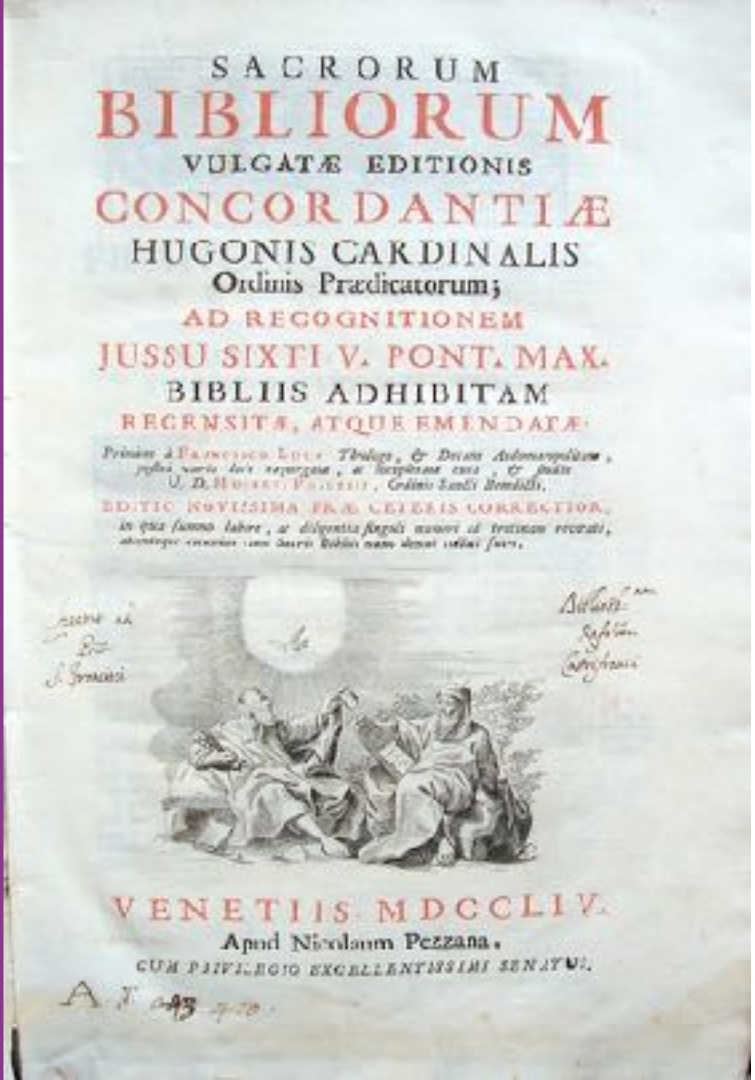
¡Hola!

Soy una lingüista que utiliza R

@rivaquiroga

si bien esta presentación es sobre palabras, partiremos con un número

1230



SACRORUM
BIBLIORUM
VULGATÆ EDITIONIS
CONCORDANTIÆ
HUGONIS CARDINALIS
Ordinis Prædicatorum;
AD RECOGNITIONEM
JUSSU SIXTI V. PONT. MAX.
BIBLIIS ADHIBITAM
RECENSITÆ, ATQUE EMENDATÆ

Primùm à FRANCISCO LUCA *Theologo, & Decano Audomaropolitano, postea variis locis repurgatæ, ac locupletatæ cura, & studio* U. D. HUBERTI PHALESII, *Ordinis Sancti Benedicti.*

EDITIO NOVISSIMA PRÆ CÆTERIS CORRECTIOR,
in qua summo labore, ac diligentia singuli numeri ad trutinam revocati, [illegible] cum Sacris Bibliis nunc denuo collati fuere.

[illegible]
S. [illegible]

[illegible]

VENETIIS MDCCLIV.
Apud Nicolaum Pezzana.
CUM PRIVILEGIO EXCELLENTISSIMI SENATUS.

A. T. 93 [illegible]

1230

Hugo de San Caro

publica una edición de la Biblia con un índice de concordancias: de cada término se ofrece el contexto en que aparece y su ubicación

SACRORUM
BIBLIORUM
VULGATÆ EDITIONIS
CONCORDANTIÆ
HUGONIS CARDINALIS
Ordinis Prædicatorum;
AD RECOGNITIONEM
JUSSU SIXTI V. PONT. MAX.
BIBLIIS ADHIBITAM
RECENSITÆ, ATQUE EMENDATÆ

Primùm à FRANCISCO LUCA *Theologo, & Decano Audomaropolitano, postea variis locis repurgatæ, ac locupletatæ curis, & studio* U. D. HUBERTI PHALESII, *Ordinis Sancti Benedicti.*

EDITIO NOVISSIMA PRÆ CETERIS CORRECTIOR,
in qua summo labore, ac diligentia singuli numeri ad textum revocati, atque [illegible] cum sacris Bibliis nunc denuo collati fuere.

Sacrae ad [illegible]
S. Ignatii

Bibliothecae
[illegible]
Capucinorum

VENETIIS MDCCLIV.
Apud Nicolaum Pezzana.
CUM PRIVILEGIO EXCELLENTISSIMI SENATUS.

A. I. 93 4.20

1230

Hugo de San Caro

(no estaba solo: un ejército de frailes lo ayudó a procesar el texto) otro número

1946

1946 Roberto Busa

con el apoyo de IBM, inicia la elaboración de un índice de concordancias de la obra de Santo Tomás: el *Index Thomisticum*

http://www.corpusthomisticum.org/it/index.age

¿lo hizo solo?

decenas de mujeres trabajaron en el proceso

imágenes: dunescholar.com

impávidas frente a la mirada masculina

imagen: dunescholar.com

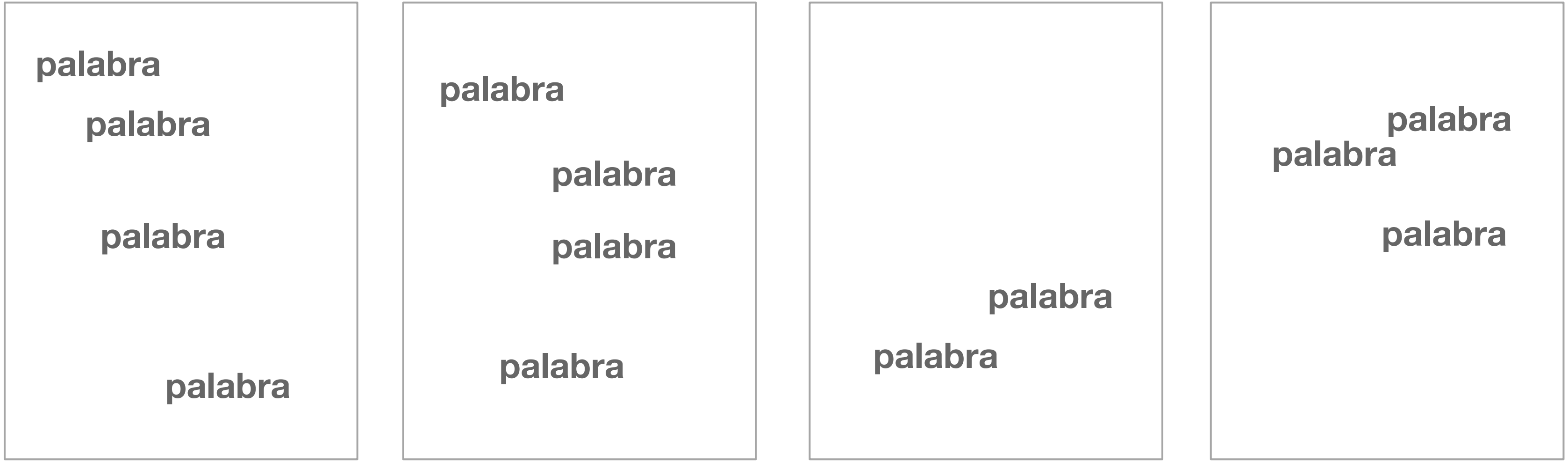
reducir información y buscar patrones

han sido algunos de los objetivos que el procesamiento y análisis de textos ha buscado ahora ya no necesitamos la ayuda de frailes: existen herramientas como R, que permiten procesar y analizar información textual.

el análisis de textos se basa en la

frecuencia y distribución

de palabras o grupos de palabras



palabra
palabra
palabra
palabra

palabra
palabra
palabra
palabra

palabra
palabra

palabra
palabra
palabra

cuánto y dónde aparecen en un texto o en un conjunto de textos (un *corpus*)

hay dos enfoques principales para abordar el análisis ignorar el orden y la función de las palabras

“bolsa de palabras”

```r
library(janeaustenr)
```

```
   word       n
   <chr>      <int>

   miss       1855
   time       1337
   dear       822
   lady       817
   sister     806
   day        797
   house      699
```

los textos se analizan a partir de tablas de frecuencia

“bolsa de palabras”

```r
library(tm)
```

```
      Terms
Docs    abarcar       abismo     bondad     buscar  ...
  1     0             0          0          0
  2     0             6          2          0
  3     0             1          0          0
  4     1             2          0          0
  5     0             1          0          1
  6     0             6          0          0
 ...
```

o matrices

“bolsa de palabras”

es lo que suele conocerse como

text mining

“bolsa de palabras”

el otro enfoque

“bolsa de palabras”

análisis y etiquetado sintáctico

para el análisis sí importa el orden y función de las palabras → ♡ gramática

♡

"bolsa de palabras"

análisis y etiquetado sintáctico

```
library(cleanNLP)

txt <- c("Hoy es la primera reunión de R-Ladies
Santiago")
```

```
id      word         upos      pos
<int>  <chr>         <chr>     <chr>
 1         Hoy       ADV       ADV___
 2         es        AUX       AUX__Mood=Ind|Number=Sing|Person=3|Tense=Pres| 3
 3         la        DET       DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art
 4         primera   ADJ       ADJ__Gender=Fem|Number=Sing|NumType=Ord
 5         reunión   NOUN      NOUN__Gender=Fem|Number=Sing
 6         de        ADP       ADP__AdpType=Prep
 7         R         PROPN     PROPN___
 8         -         PUNCT     PUNCT__PunctType=Dash
 9         Ladies    PROPN     PROPN___
10         Santiago  PROPN     PROPN___
```

“bolsa de palabras”

análisis y etiquetado sintáctico

es lo que suele hacerse desde la

lingüística

"bolsa de palabras"

análisis y etiquetado sintáctico

En esta presentación mostraré

qué se puede hacer en R

(no *cómo*, porque solo tenemos 10 minutos)

y algunos **referentes**

para que se entusiasmen y sigan explorando

O'REILLY

Text Mining with R

A TIDY APPROACH

Julia Silge & David Robinson

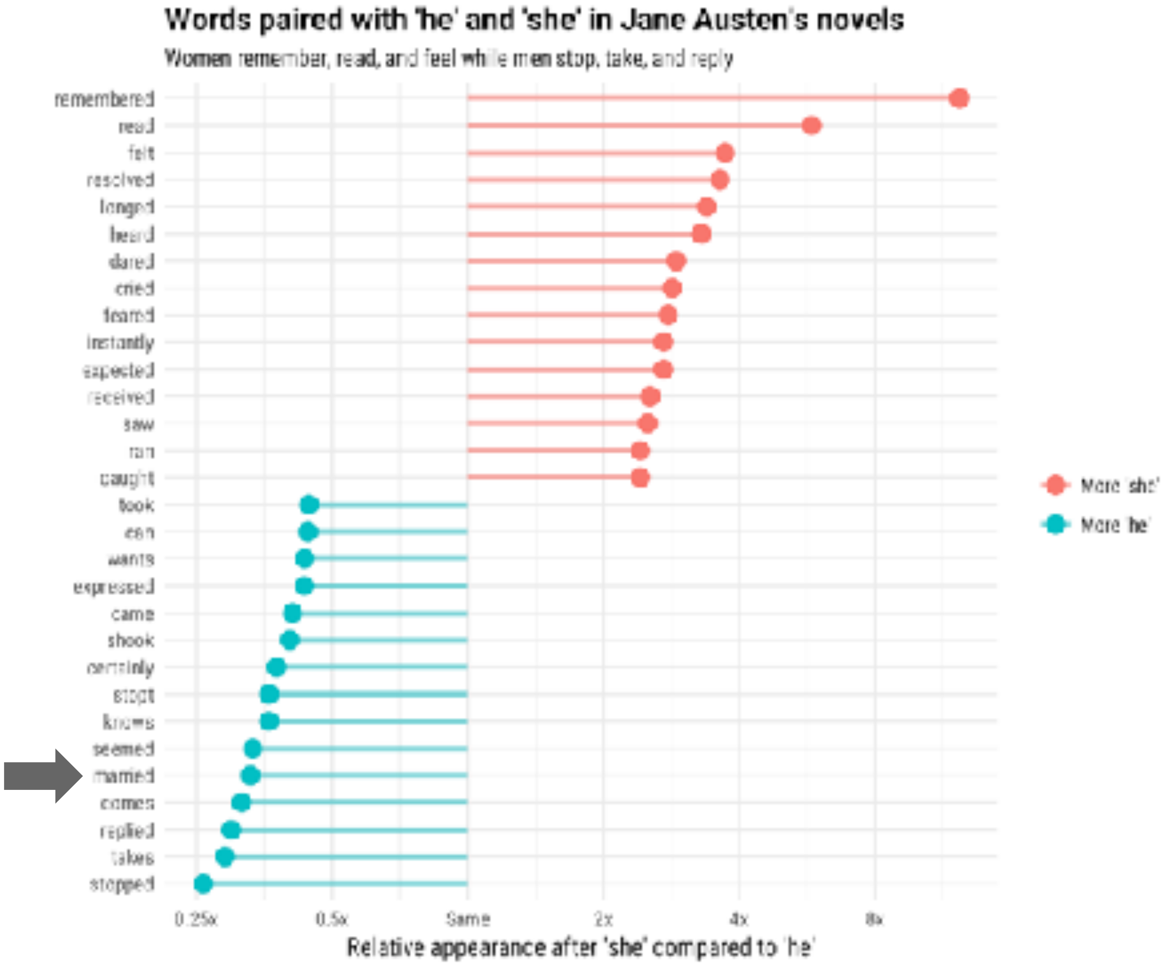
`library(tidytext)`

http://tidytextmining.com

JULIA SILGE

@jualiasilge

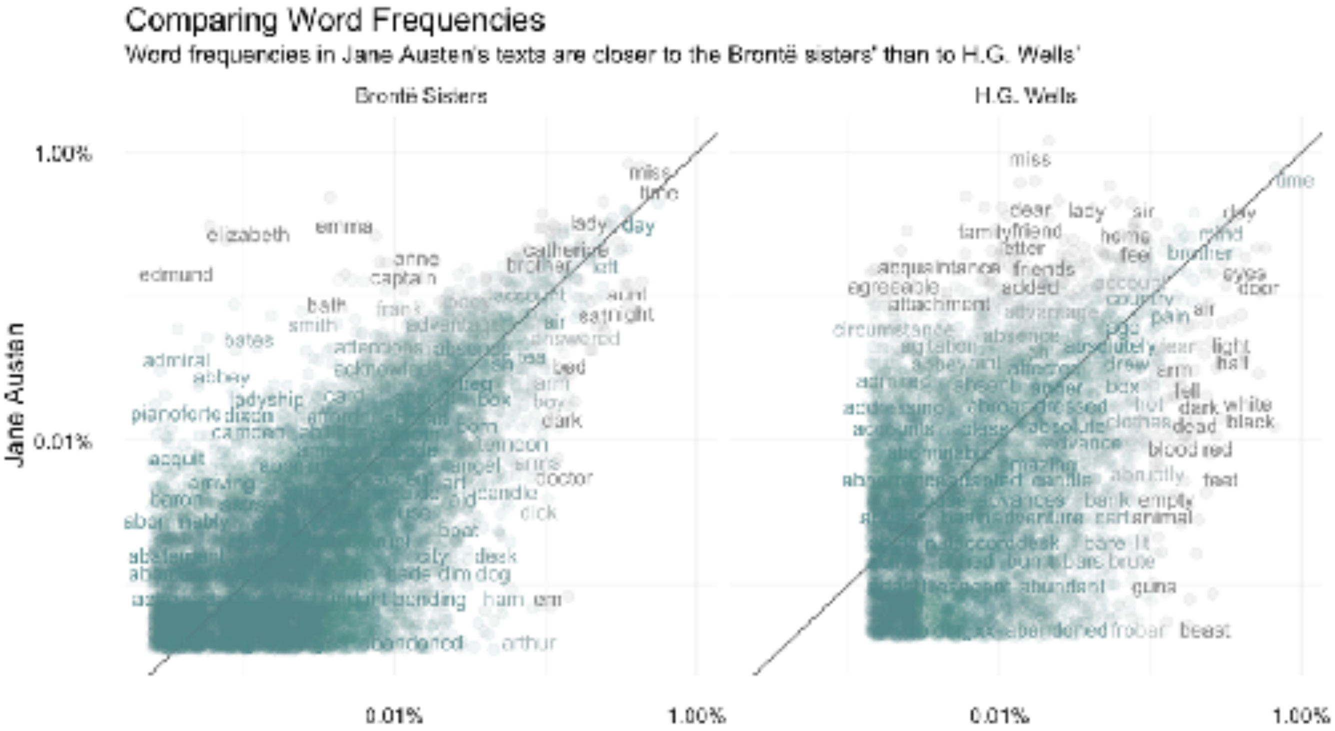
juliasilge.com



JULIA SILGE

@jualiasilge

juliasilge.com



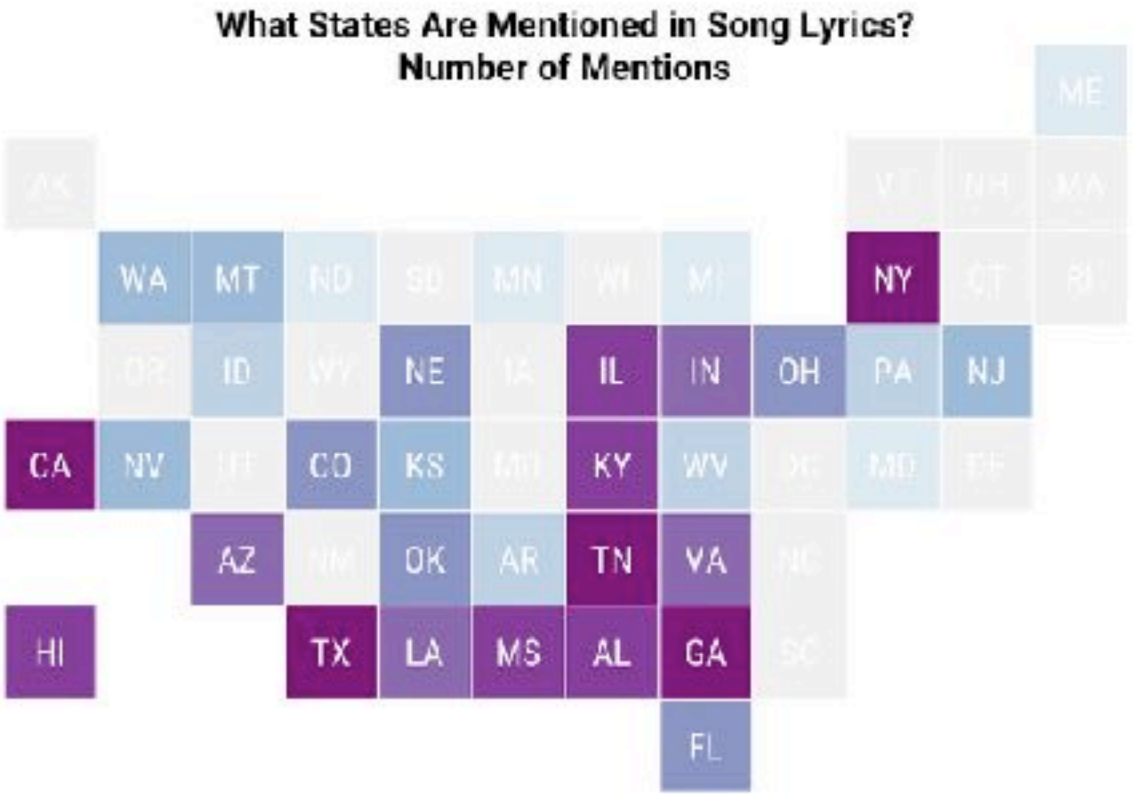
Comparing Word Frequencies

Word frequencies in Jane Austen's texts are closer to the Brontë sisters' than to H.G. Wells'

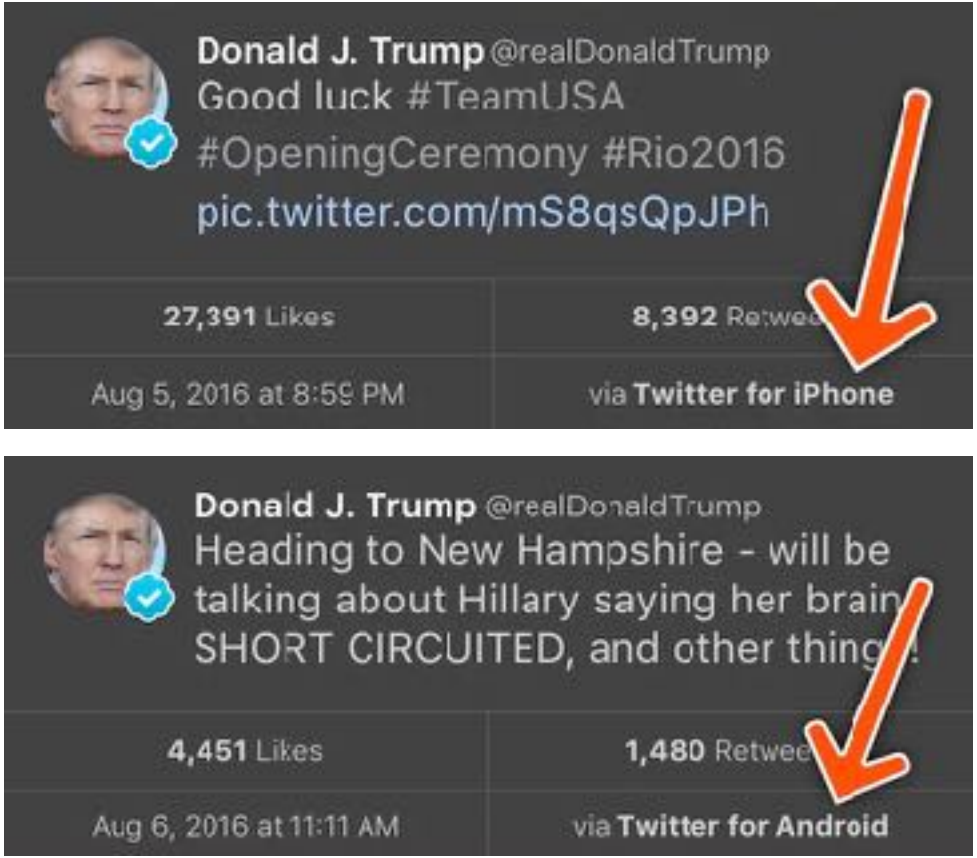
JULIA SILGE

@jualiasilge

juliasilge.com



What States Are Mentioned in Song Lyrics?
Number of Mentions



Donald J. Trump @realDonaldTrump
Good luck #TeamUSA
#OpeningCeremony #Rio2016
pic.twitter.com/mS8qsQpJPh

27,391 Likes | 8,392 Retwee

Aug 5, 2016 at 8:59 PM | via Twitter for iPhone

Donald J. Trump @realDonaldTrump
Heading to New Hampshire - will be talking about Hillary saying her brain SHORT CIRCUITED, and other things!

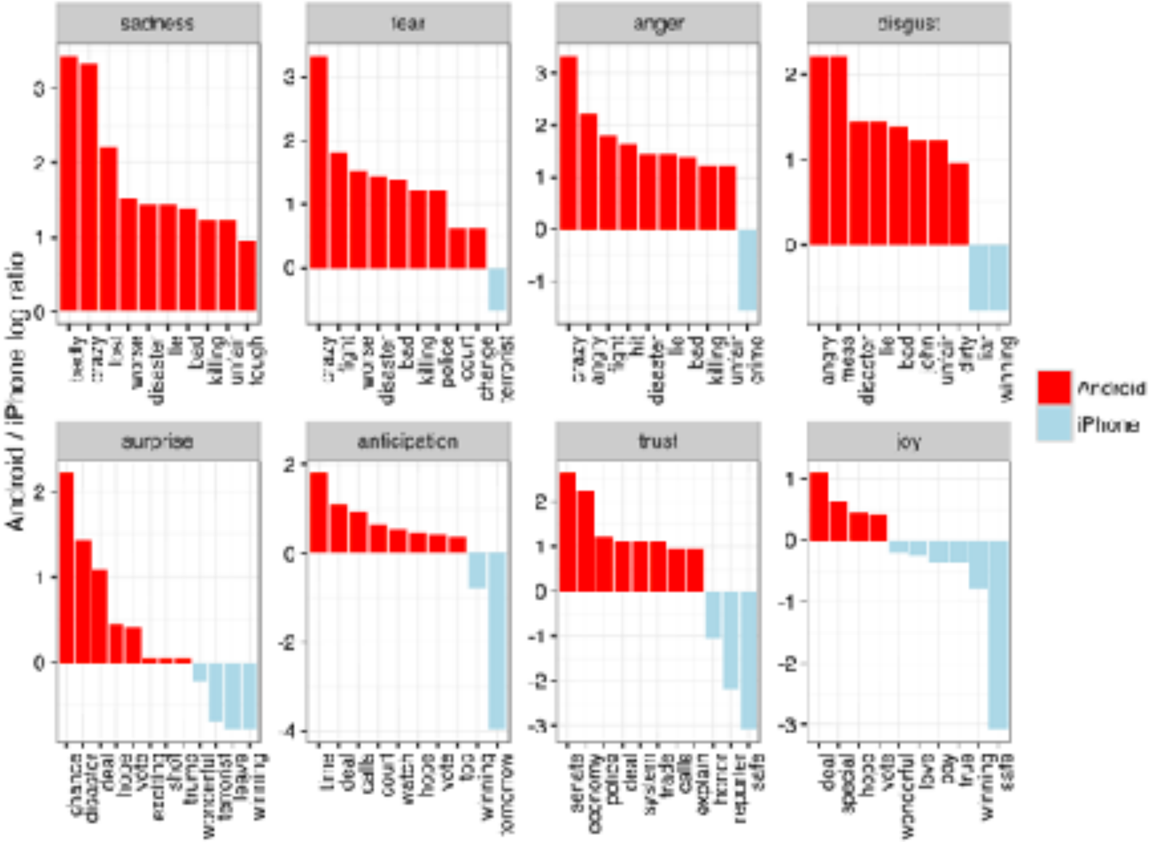
4,451 Likes | 1,480 Retwee

Aug 6, 2016 at 11:11 AM | via Twitter for Android

Trump escribe la mitad enojada de sus tweets (los de Android)

David
Robinson

@drob
varianceexplained.org



David Robinson

@drob

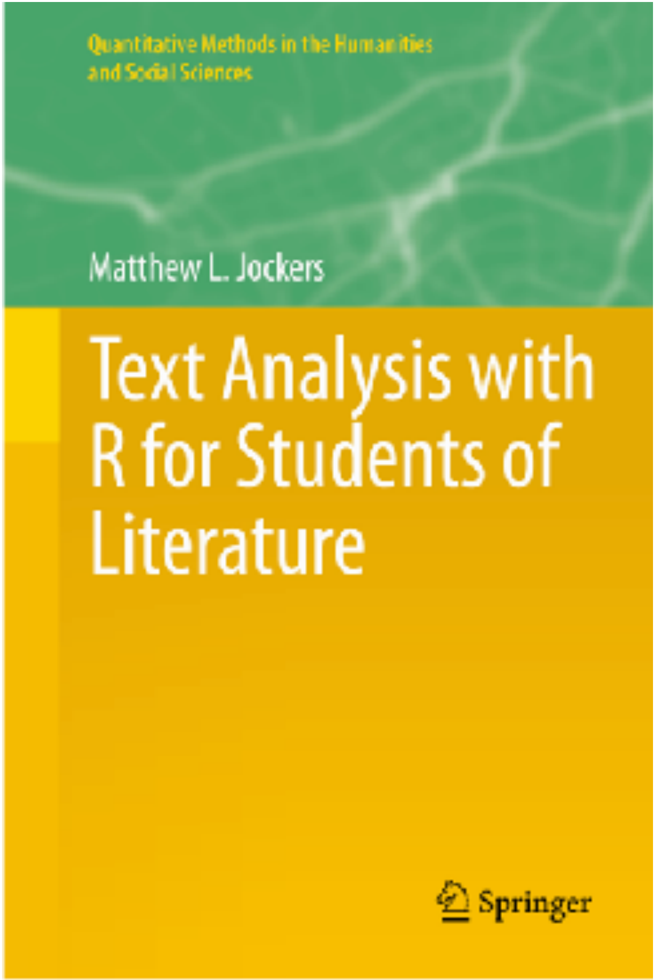
varianceexplained.org

análisis de sentimientos a través de tidytext

Matthew L. Jockers

@mljockers

matthewjockers.net



Quantitative Methods in the Humanities and Social Sciences

Matthew L. Jockers

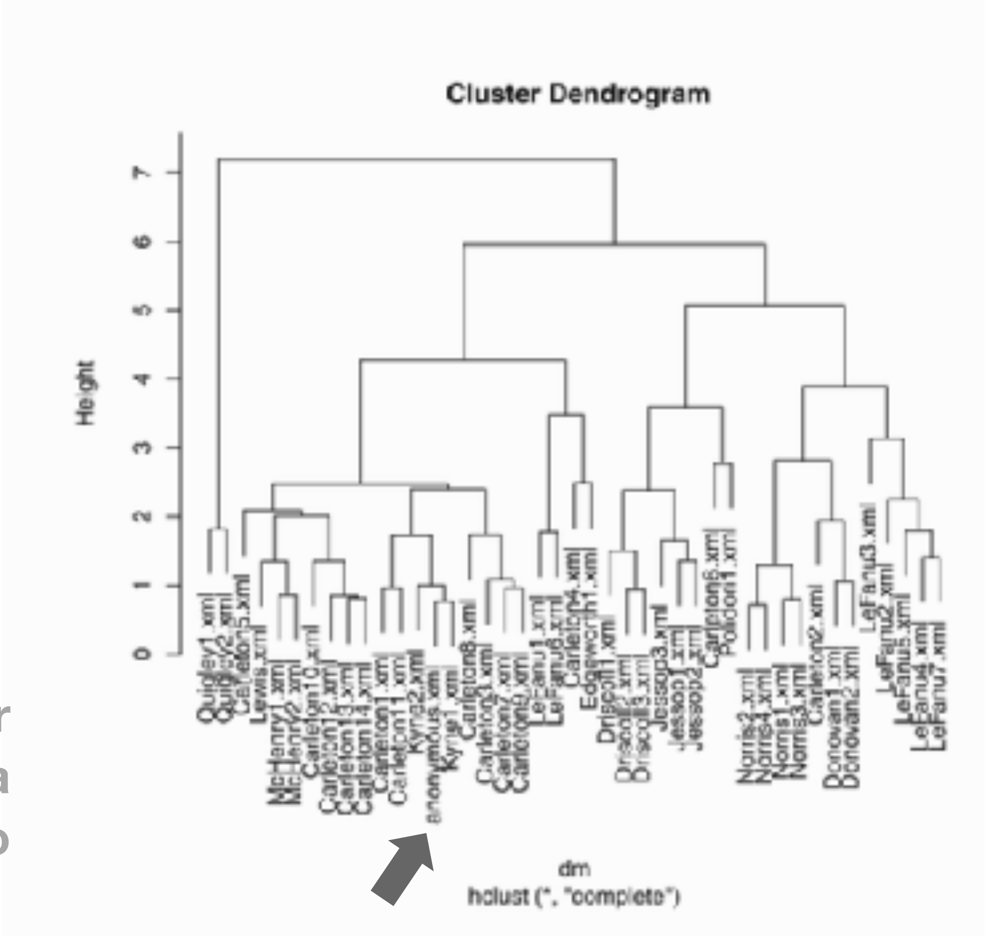
Text Analysis with R for Students of Literature

Springer

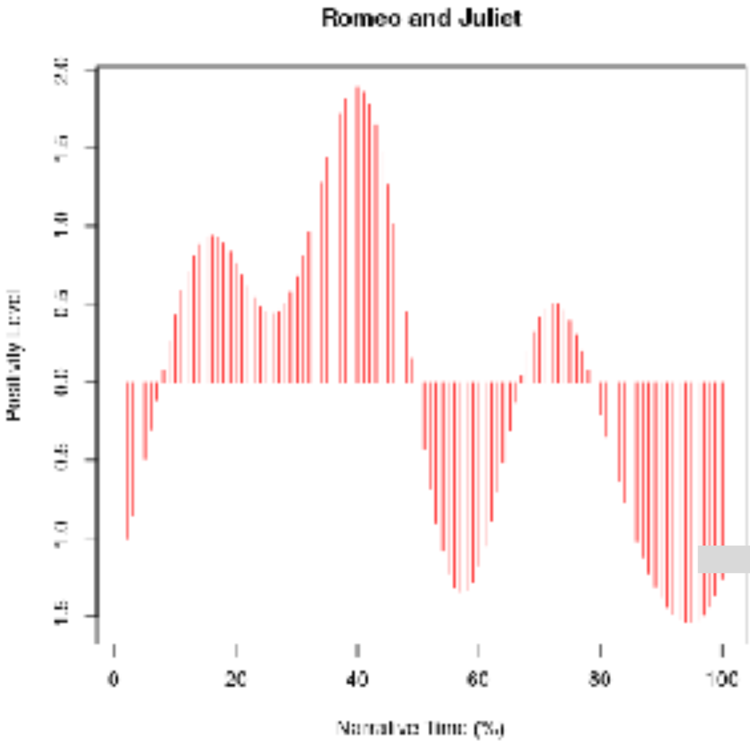
Matthew L. Jockers

@mljockers

matthewjockers.net



análisis estilométrico: por ejemplo, para atribuir la autoría de un texto



```r
library(syuzhet)
```

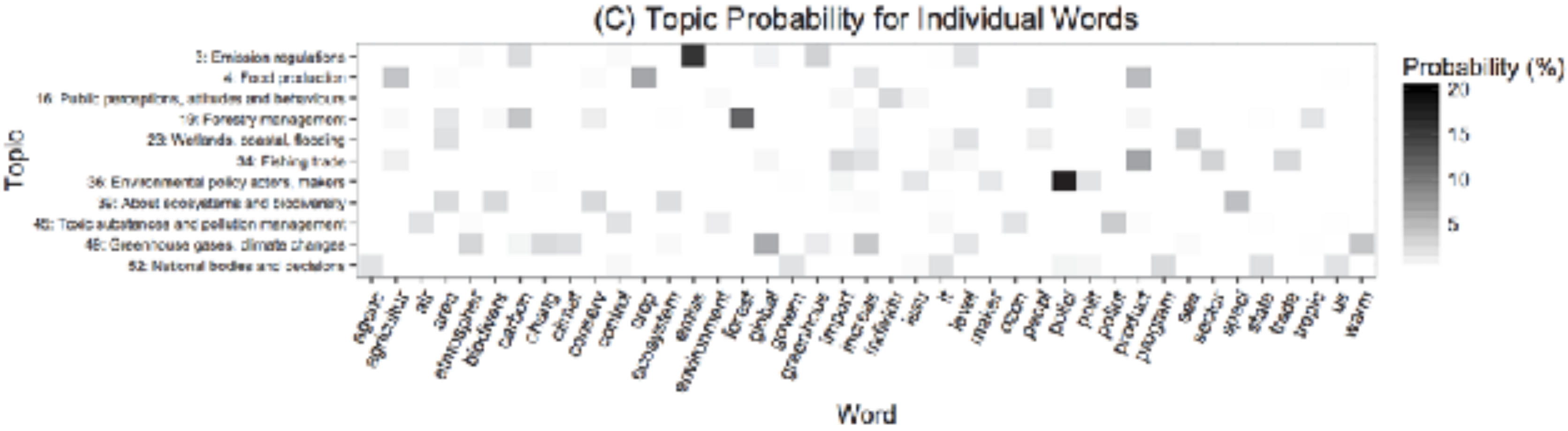
análisis de sentimientos + movimientos en la trama

todos mueren :(

fuente: http://projectalexandria.net/

```
library(topicmodels)
```

modelación de tópicos

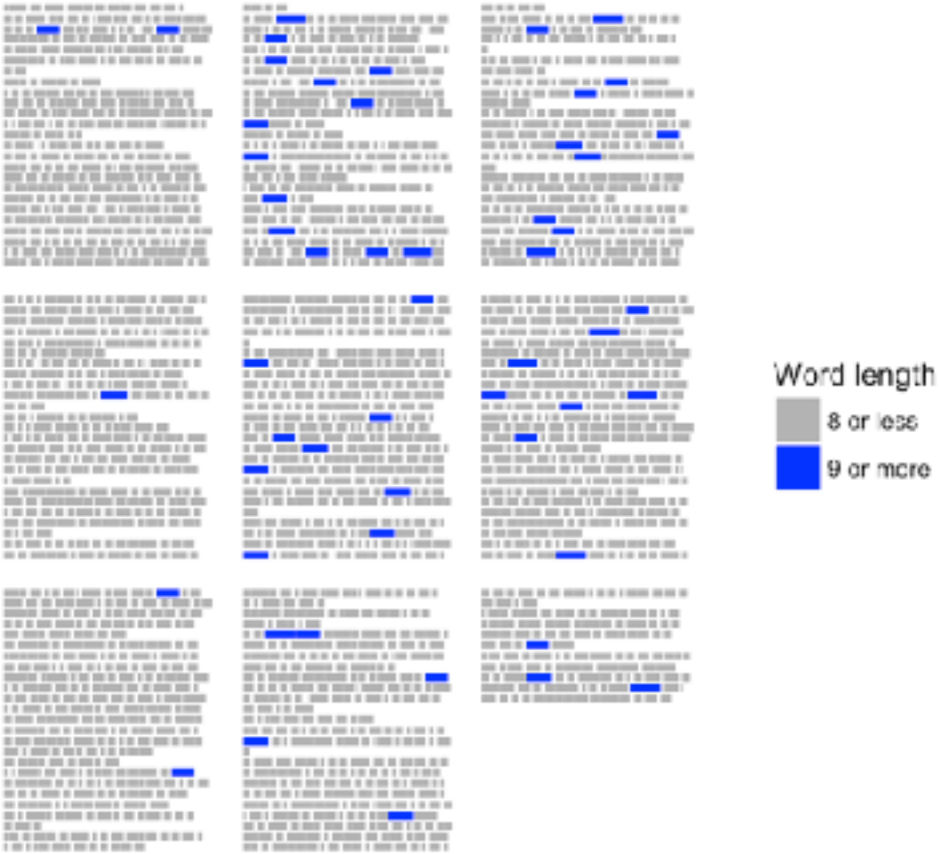


Murakami et al. (2017). ‘What is this corpus about?’: using topic modelling to explore a specialised corpus. *Corpora 12*(2), 243 – 277.

```
library(tm)
library(mallet)
```

library(ggpage)

@Emil_Hvitfeldt



también podemos visualizar distribuciones

library(cleanNLP)

Taylor Arnold @statsmaths

```
txt <- c("se levantó una nación noble, valiente y
solidaria")
id      word        upos      pos
<int>  <chr>       <chr>     <chr>
1        se        PRON      PRON__Person=3
2   levantó        VERB      VERB__Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
3       una        DET       DET__Definite=Ind|Gender=Fem|Number=Sing|PronType=Art
4    nación        NOUN      NOUN__Gender=Fem|Number=Sing
5     noble        ADJ       ADJ__Number=Sing
6         ,        PUNCT     PUNCT__PunctType=Comm
7  valiente        ADJ       ADJ__Number=Sing
8         y        CONJ      CCONJ___
9 solidaria        ADJ       ADJ__Gender=Fem|Number=Sing
```

¿qué candidato chileno a la presidencia tiene fama de acumular adjetivos? Con este análisis podríamos identificarlo

```
library(quanteda)
```

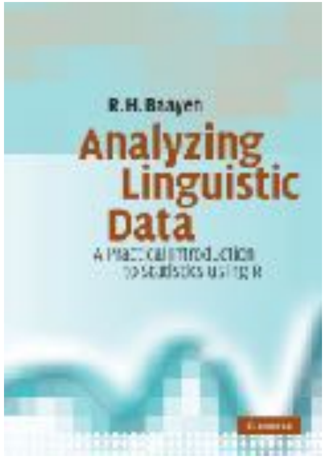
Kenneth Benoit @kenbenoit

quanteda

Quantitative Analysis of Textual Data

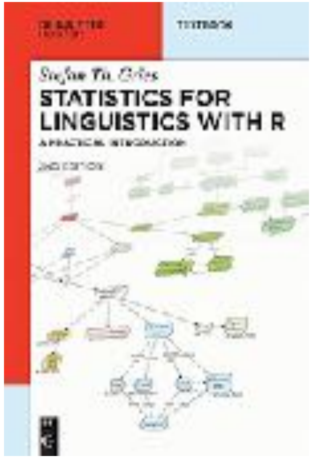
Paquete que permite realizar análisis propios de la lingüística de corpus:

```
kwic( )
keyness( )
```

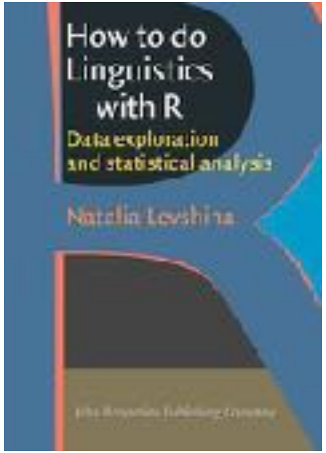

R.H. Baayen
Analyzing Linguistic Data
A Practical Introduction to Statistics using R

(Baayen, 2008)

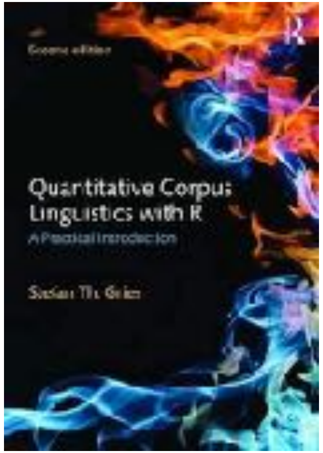
(Johnson, 2008)



(Gries, 2013)

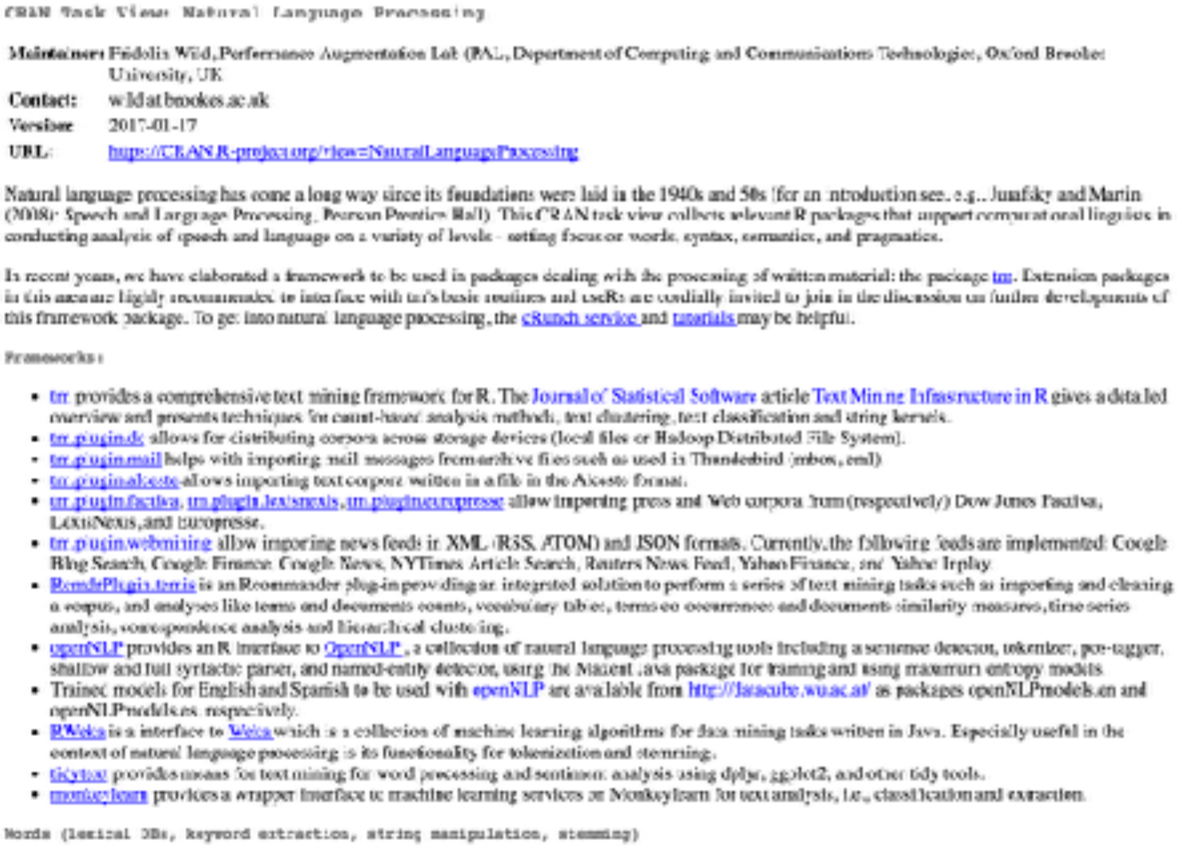

How to do Linguistics with R
Data exploration and statistical analysis
Natalia Levshina

(Levshina, 2015)



(Gries, 2016)

para las lingüistas que quieren animarse a utilizar R https://cran.r-project.org/



CRAN Task View: Natural Language Processing

Maintainer:	Fridolin Wild, Performance Augmentation Lab (PAL), Department of Computing and Communications Technologies, Oxford Brookes University, UK
Contact:	wild at brookes.ac.uk
Version:	2017-01-17
URL:	https://CRAN.R-project.org/view=NaturalLanguageProcessing

Natural language processing has come a long way since its foundations were laid in the 1940s and 50s (for an introduction see, e.g., Jurafsky and Martin (2008): Speech and Language Processing, Pearson Prentice Hall). This CRAN task view collects relevant R packages that support computational linguists in conducting analysis of speech and language on a variety of levels - setting focus on words, syntax, semantics, and pragmatics.

In recent years, we have elaborated a framework to be used in packages dealing with the processing of written material: the package tm. Extension packages in this area are highly recommended to interface with tm's basic routines and useRs are cordially invited to join in the discussion on further developments of this framework package. To get into natural language processing, the cRunch service and tutorials may be helpful.

Frameworks:

- tm provides a comprehensive text mining framework for R. The Journal of Statistical Software article Text Mining Infrastructure in R gives a detailed overview and presents techniques for count-based analysis methods, text clustering, text classification and string kernels.
- tm.plugin.dc allows for distributing corpora across storage devices (local files or Hadoop Distributed File System).
- tm.plugin.mail helps with importing mail messages from archive files such as used in Thunderbird (mbox, eml).
- tm.plugin.alceste allows importing text corpora written in a file in the Alceste format.
- tm.plugin.factiva, tm.plugin.lexisnexis, tm.plugin.europresse allow importing press and Web corpora from (respectively) Dow Jones Factiva, LexisNexis, and Europresse.
- tm.plugin.webmining allow importing news feeds in XML (RSS, ATOM) and JSON formats. Currently, the following feeds are implemented: Google Blog Search, Google Finance, Google News, NYTimes Article Search, Reuters News Feed, Yahoo Finance, and Yahoo Inplay.
- RcmdrPlugin.temis is an Rcommander plug-in providing an integrated solution to perform a series of text mining tasks such as importing and cleaning a corpus, and analyses like terms and documents counts, vocabulary tables, terms co-occurrences and documents similarity measures, time series analysis, correspondence analysis and hierarchical clustering.
- openNLP provides an R interface to OpenNLP, a collection of natural language processing tools including a sentence detector, tokenizer, pos-tagger, shallow and full syntactic parser, and named-entity detector, using the Maxent Java package for training and using maximum entropy models.
- Trained models for English and Spanish to be used with openNLP are available from http://datacube.wu.ac.at/ as packages openNLPmodels.en and openNLPmodels.es, respectively.
- RWeka is a interface to Weka which is a collection of machine learning algorithms for data mining tasks written in Java. Especially useful in the context of natural language processing is its functionality for tokenization and stemming.
- tidytext provides means for text mining for word processing and sentiment analysis using dplyr, ggplot2, and other tidy tools.
- monkeylearn provides a wrapper interface to machine learning services on Monkeylearn for text analysis, i.e., classification and extraction.

Words (lexical DBs, keyword extraction, string manipulation, stemming)

¡Hay muchos más paquetes! Ver en Task View > Natural Language Processing