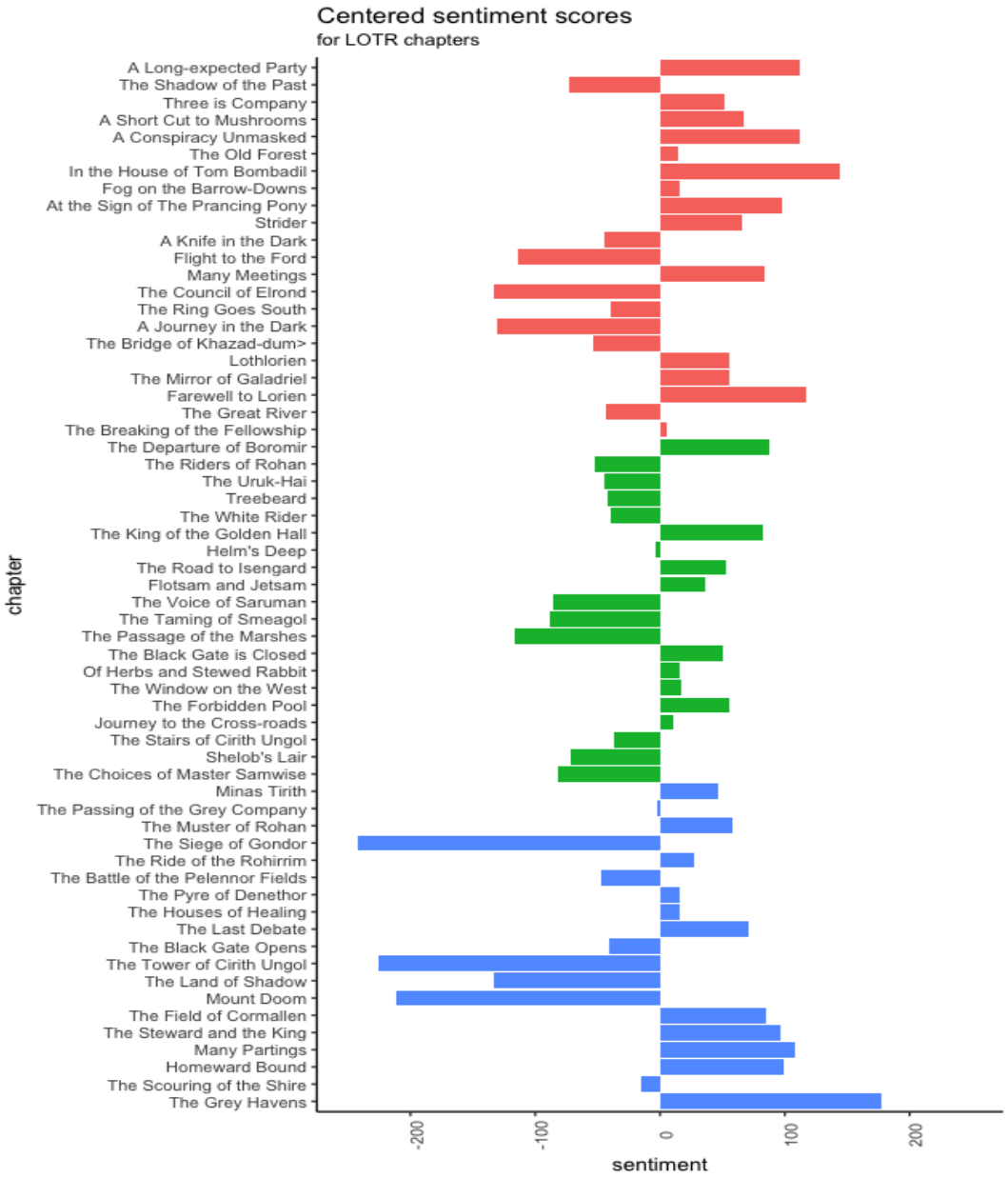
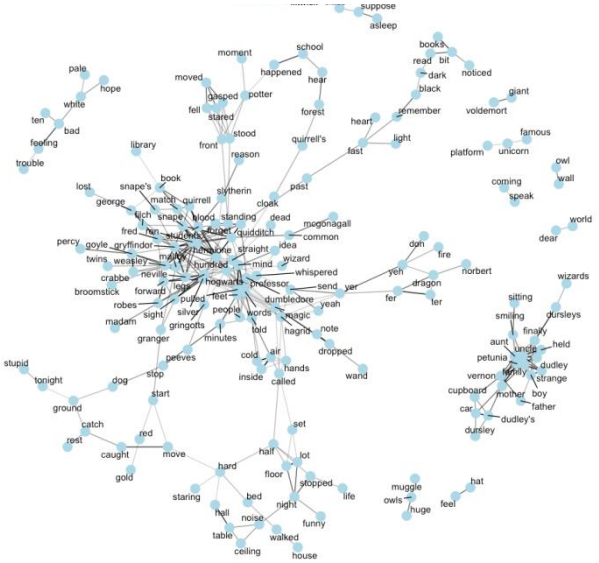
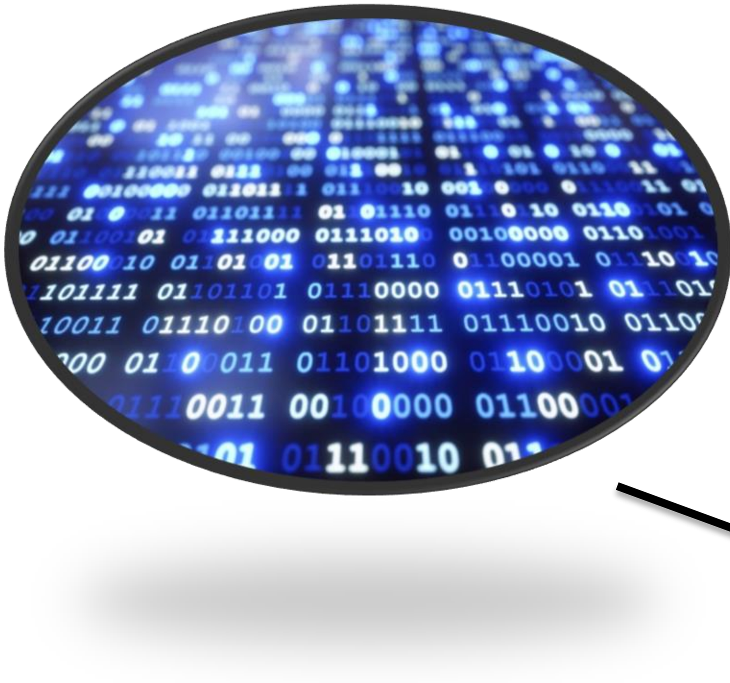
Minería de texto





Oscar Centeno Mora

Nuevo tipo de análisis



$$\Xi_c = (0.136,\ 0.307,\ 0.057^+ \ \|\ 0.194^-,\ 0.148,\ 0.087, 0.074) \ \text{ and } \ \aleph_c = [\varnothing].$$

A lot of text goes here. A lot of text goes here.

$$\Xi_c = (0.136,\ 0.307,\ 0^+ \ \|\ 0.137^-,\ 0.148,\ 0.087, 0.074) \ \text{ and } \ \aleph_c = [\overset{0.057}{(CC)}].$$

Nuevo tipo de análisis

- Hasta ahora hemos analizado únicamente números y más números…

- El presente capítulo presenta una forma de analizar texto.

- Este tipo de análisis se conoce como minería de texto, y es muy útil para explorar los datos y también encontrar ciertos patrones.

- Recordemos que el presente curso trata los temas de forma descriptiva, por lo que se centra en el análisis descriptivo de los datos.


enterprise
infrastructure
technology
operations
cards
text mining
objectives
solution
stakeholder

Índice

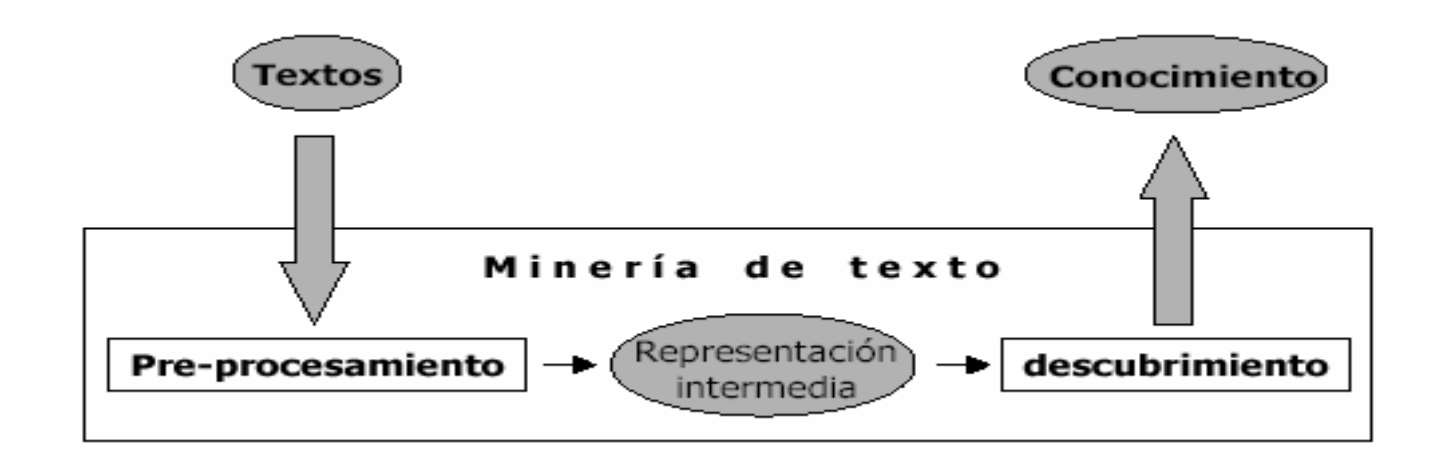
Índice

Índice

Introducción

- La minería de texto, también conocida como minería de datos de texto, aproximadamente equivalente a la analítica de texto, es el proceso de derivar información de alta calidad a partir del texto.

- La información se deriva típicamente a través del diseño de patrones y tendencias a través de medios tales como el aprendizaje de patrones estadísticos



Introducción

- La minería de textos generalmente implica el proceso de estructurar el texto de entrada (generalmente el análisis, junto con la adición de algunas características lingüísticas derivadas y la eliminación de otras, y su posterior inserción en una base de datos), derivando patrones dentro de los datos estructurados y, finalmente, evaluación e interpretación de la salida.
- Las tareas típicas de minería de texto incluyen la categorización del texto, la agrupación de texto, la extracción de conceptos / entidades, la producción de taxonomías granulares, el análisis de sentimientos, el resumen de documentos y el modelado de relaciones de entidades (es decir, las relaciones de aprendizaje entre entidades nombradas)

Introducción

Text Mining

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

Introducción

Minería de Texto

Tipo de estudio

- Relación entre proteínas
- Anotaciones en el documento
- Identificación de secuencia entre proteínas
- Recuperación de texto
- Agrupamiento de documentos por temática
- Identificación de entidades
- Identificación de asociaciones entre genes y enfermedades
- Extracción de acrónimos y de términos
- Etiquetado de entidades biológicas

Recuperación de la información

Extracción de la información

Procesamiento de información

Métodos de minería

Agrupamiento
Clasificación
Resumen

Identificación de términos y relaciones

Análisis de información y visualización

Introducción

- El análisis de texto implica la recuperación de información, el análisis léxico para estudiar las distribuciones de frecuencia de palabras, el reconocimiento de patrones, el etiquetado / anotación, la extracción de información, las técnicas de extracción de datos, incluido el análisis de enlaces y asociaciones, la visualización y el análisis predictivo.

- El objetivo general es, esencialmente, convertir el texto en datos para el análisis, a través de la aplicación del procesamiento del lenguaje natural (PNL) y los métodos analíticos.

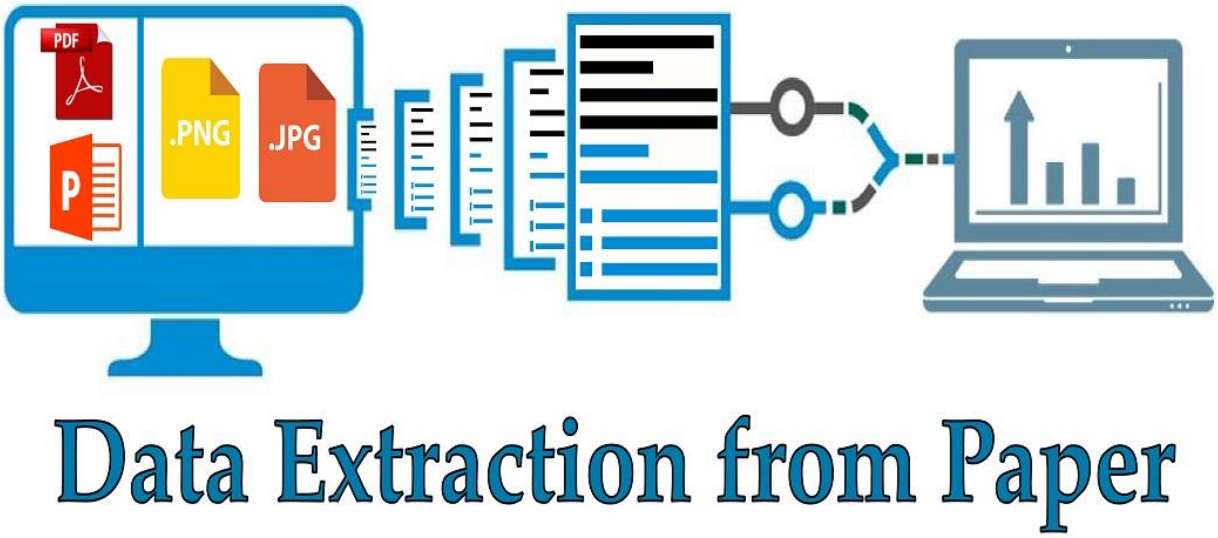
Índice

Fuentes de los datos

¿De dónde podemos obtener los datos?

Fuentes de los datos

En la extracción de datos para llevar a cabo el análisis de texto, podemos obtenerlo a partir de:



-Periódicos
-Libros
-Informes
-Cartas
-Páginas web
-Tweets, Facebook
-Y cualquier forma donde este se exponga.

Todas las formas requieren un cierta forma de extraerla, pero en este curso veremos los más conocidos.

Índice

Limpieza, tokenización, Corpus y matriz de términos

El proceso de limpieza de texto, dentro del ámbito de *text mining*, consiste en eliminar del texto todo aquello que no aporte información sobre su temática, estructura o contenido, además de estandarizar las palabras para su previo análisis.

No existe una única forma de hacerlo, depende en gran medida de la finalidad del análisis y de la fuente de la que proceda el texto. Por ejemplo, en las redes sociales los usuarios pueden escribir de la forma que quieran, lo que suele resultar en un uso elevado de abreviaturas y signos de puntuación.

En la limpieza de datos se podría proceder a eliminar o estandarizar:

- Patrones no informativos (*urls* de páginas web)
- Signos de puntuación
- Mayústuculas / minúsculas
- Signos de puntuación
- Etiquetas HTML
- Caracteres sueltos
- Números, etc…

Limpieza, tokenización, Corpus y matriz de términos

Algunas funciones en R que ayudan en el proceso de limpieza de datsos:

Form base R:

- `tolower`
  - helpful for term aggregation but can be harmful if you are trying to identify proper nouns like cities

From tm package:

- `removePunctiation`
  - very helpful with social media, but harmful if you are trying to identify emoticons with punctuation like smiley faces
- `removeNumbers`
  - don't do this if you currencies or quantities
- `stripWhiteSpace`
  - sometimes there is extra lines or whitespace
- `removeWords`
  - usually words like 'the' and 'of' are not very helpful

Limpieza, tokenización, Corpus y matriz de términos

Other useful tm functions

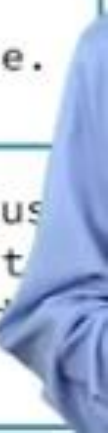
- `tm_map` takes a corpus and a processing funciton and transforms the corpus
  - if the function is not from the tm library than wrap it in `content_transformer()` function
- `stemDocument` functions can also be very helpful
  - but you often have tokens that are not words
  - use stem Completion to turn the stems back into words

From the qdap package

- `bracketX()` : Remove all text within brackets (e.g. "It's (so) cool" becomes "It's cool")
- `replace_number()` : Replace numbers with their word equivalents (e.g. "2" becomes "two")
- `replace_abbreviation()` : Replace abbreviations with their full text equivalents (e.g. "Sr" becomes "Senior")
- `replace_contraction()` : Convert contractions back to their base words (e.g. "shouldn't" becomes "should not")
- `replace_symbol()` : Replace common symbols with their word equivalents (e.g. "$" becomes "dollar")

Limpieza, tokenización, Corpus y matriz de términos

TM Function	Description	Before	After
`tolower()`	Makes all text lowercase	`Starbucks is from Seattle.`	`starbucks is from seattle.`
`removePunctuation()`	Removes punctuation like periods and exclamation points	`Watch out! That coffee is going to spill!`	`Watch out That coffee is going to spill`
`removeNumbers()`	Removes numbers	`I drank 4 cups of coffee 2 days`	`I drank cups of coffee days ago.`
`stripWhiteSpace()`	Removes tabs and extra spaces	`I like coffee.`	`I like coffee.`
`removeWords()`	Removes specific words (e.g. "the", "of") defined by the data scientist	`The coffee house and barista he visited were nice, she said hello.`	`The coffee hous barista visit nice, sai hello.`



Limpieza, tokenización, Corpus y matriz de términos

- **Tokenizar** un texto consiste en dividir el texto en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso, las palabras.

- Existen múltiples librerías que automatizan en gran medida la limpieza y tokenización de texto, por ejemplo, tokenizers o quanteda. Sin embargo, se entiende mejor el proceso implemento una función propia que, si bien puede estar menos optimizada, es más transparente. Definir una función que contenga cada uno de los pasos de limpieza tiene la ventaja de poder adaptarse fácilmente dependiendo del tipo de texto analizado.

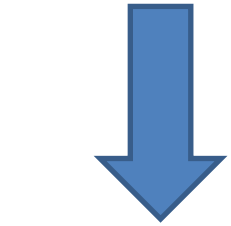
Limpieza, tokenización, Corpus y matriz de términos

Veamos un caso de tokenización. Primero creamos una función, luego a partir de un texto se trae lo que se va a analizar, y le brindamos a cada palabra una especificación de “elemento”, para su próximo análisis.

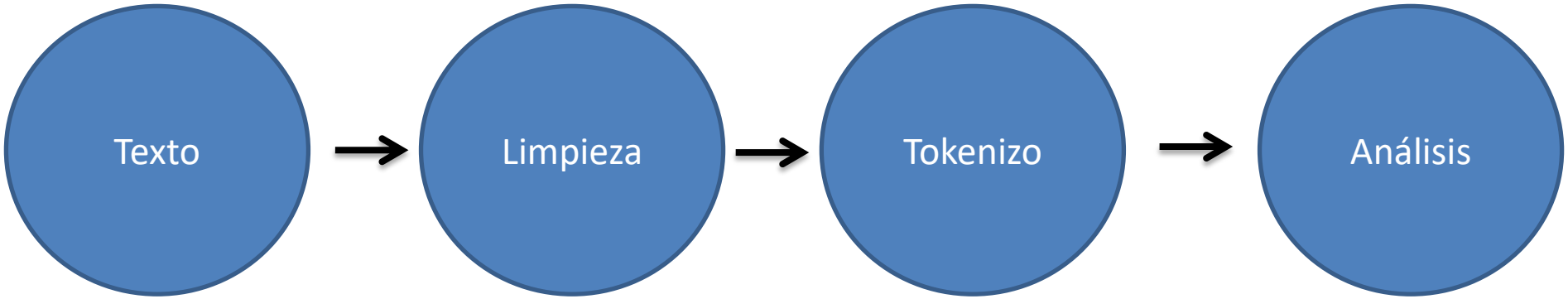
```
limpiar_tokenizar <- function(texto){
  # El orden de la limpieza no es arbitrario
    # Se convierte todo el texto a minúsculas
    nuevo_texto <- tolower(texto)
    # Eliminación de páginas web (palabras que empiezan por "http." seguidas
    # de cualquier cosa que no sea un espacio)
    nuevo_texto <- str_replace_all(nuevo_texto,"http\\S*", "")
    # Eliminación de signos de puntuación
    nuevo_texto <- str_replace_all(nuevo_texto,"[[:punct:]]", " ")
    # Eliminación de números
    nuevo_texto <- str_replace_all(nuevo_texto,"[[:digit:]]", " ")
    # Eliminación de espacios en blanco múltiples
    nuevo_texto <- str_replace_all(nuevo_texto,"[\\s]+", " ")
    # Tokenización por palabras individuales
    nuevo_texto <- str_split(nuevo_texto, " ")[[1]]
    # Eliminación de tokens con una longitud < 2
    nuevo_texto <- keep(.x = nuevo_texto, .p = function(x){str_length(x) > 1})
    return(nuevo_texto)
}
```

```
test = "Esto es 1 ejemplo de l'limpieza de6 TEXTO  https://t.co/rnHPgyhx4Z @JoaquinAmatRodrigo #text
mining"
limpiar_tokenizar(texto = test)
```

```
## [1] "esto"             "es"               "ejemplo"
## [4] "de"               "limpieza"         "de"
## [7] "texto"            "joaquinamatrodrigo" "textmining"
```



Limpieza, tokenización, Corpus y matriz de términos



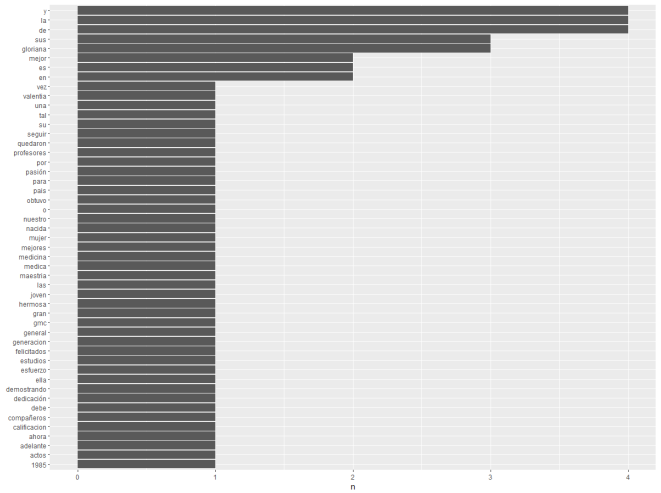
Limpieza, tokenización, Corpus y matriz de términos

- Otra forma de tokenizar podría ser utilizando la función “unnest_tokens()” de la librería tidytxt.

- Esta hace algo similar: pasar un texto normal, a un conjunto de palabras para su análisis.

```
" Gloriana, mujer joven y hermosa nacida en 1985 es una -",
"de las mejores o tal vez la mejor medica general de nuestro pais",
"EN sus estudios de maestria obtuvo la mejor calificacion de la generacion -",
" y sus actos quedaron felicitados por sus profesores y compañeros",
 "Ahora Gloriana debe seguir adelante demostrando su gran valentia, -",
  "esfuerzo, pasión y dedicación para la medicina. Ella es Gloriana, GMC")
```

```
# A tibble: 65 x 2
    line word
   <int> <chr>
 1     1 gloriana
 2     1 mujer
 3     1 joven
 4     1 y
 5     1 hermosa
 6     1 nacida
 7     1 en
 8     1 1985
 9     1 es
10     1 una
# ... with 55 more rows
```



Limpieza, tokenización, Corpus y matriz de términos

En lingüística, un corpus (corpus plural) o corpus de texto es un conjunto amplio y estructurado de textos (hoy en día generalmente almacenados y procesados electrónicamente). En la lingüística de corpus, se utilizan para realizar análisis estadísticos y pruebas de hipótesis, verificar ocurrencias o validar reglas lingüísticas dentro de un territorio lingüístico específico.

Una vez se haya preparado el documento (limpieza y tokenización), se procede a crear el Corpus, es decir, el acervo del documento a analizar.

De forma de análisis, es un primer tipo de almacenamiento de la información o el texto.

Limpieza, tokenización, Corpus y matriz de términos

Veamos un ejemplo de Corpus

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1580
##
##      [1]



##      [2]
```

```
##  [791] niebla i  aparecer augusto  puerta  casa extendi<U+663C><U+3E33>  brazo derecho  mano palma abajo  abierta  d
irigiendo  ojos  cielo
```

```
##  [792] <U+393C><U+3E37>pues  trata  mujeres  apellido deb<U+653C><U+3E64>a cambiarse  dominga  si  <U+623C><U+3E66>d<U
+663C><U+3E33>nde est<U+653C><U+3E31>  concordancia <U+393C><U+3E37>no  conozco se<U+663C><U+3E31>or <U+393C><U+3E37>y d<U
+653C><U+3E64>game d<U+653C><U+3E64>game<U+393C><U+3E37>sin sacar  dedos  bolsillo<U+393C><U+3E37> <U+623C><U+3E66>c<U+663
C><U+3E33>mo  sale as<U+653C><U+3E64> sola <U+623C><U+3E66>es soltera  casada <U+623C><U+3E66>tiene padres <U+393C><U+3E3
7>es soltera  hu<U+653C><U+3E39>rfana vive  t<U+653C><U+3E64>os

##  [793] espina  camino  vida<U+623C><U+3E62> <U+613C><U+3E31>si  aqu<U+653C><U+3E64>  hacer florecer  rosa  primera
espina <U+613C><U+3E62>si viviera  madre encontrar<U+653C><U+3E64>a soluci<U+663C><U+3E33>n  esto<U+393C><U+3E37>se dijo a
ugusto<U+393C><U+3E37>  despu<U+653C><U+3E39>s  m<U+653C><U+3E31>s dif<U+653C><U+3E64>cil  ecuaci<U+663C><U+3E33>n  se
gundo grado  fondo m<U+653C><U+3E31>s  ecuaci<U+663C><U+3E33>n  segundo grado<U+623C><U+3E62>  d<U+653C><U+3E39>biles
 quejidos  pobre animal interrumpieron  soliloquio escudri<U+663C><U+3E31><U+663C><U+3E33>  ojos  acab<U+663C><U+3E33>
 descubrir  verdura  matorral  pobre cachorrillo  perro  parec<U+653C><U+3E64>a
##  [794] buscar camino  tierra <U+613C><U+3E62><U+613C><U+3E31>pobrecillo<U+393C><U+3E37>se dijo<U+393C><U+3E37>  dejado
 reci<U+653C><U+3E39>n nacido  muera  falt<U+663C><U+3E33> valor  matarlo<U+623C><U+3E62>  recogi<U+663C><U+3E33>  anima
```

Limpieza, tokenización, Corpus y matriz de términos

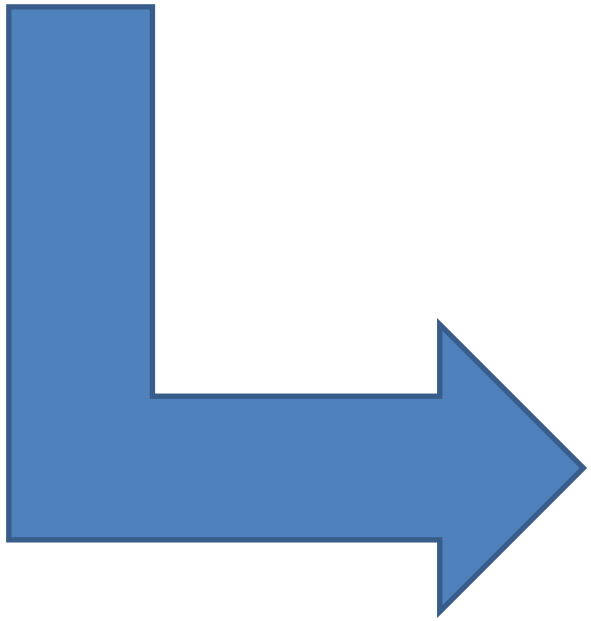
Una matriz de término de documento o matriz de documento de término es una matriz matemática que describe la frecuencia de los términos que aparecen en una colección de documentos.

En una matriz de términos de documentos, las filas corresponden a los documentos de la colección y las columnas corresponden a los términos. Hay varios esquemas para determinar el valor que cada entrada en la matriz debe tomar. Uno de esos esquemas es tf-idf. Son útiles en el campo del procesamiento del lenguaje natural.

Al final lo veo como una simple matriz pero con observaciones que son las variables y se contabiliza la frecuencias.

Limpieza, tokenización, Corpus y matriz de términos

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```



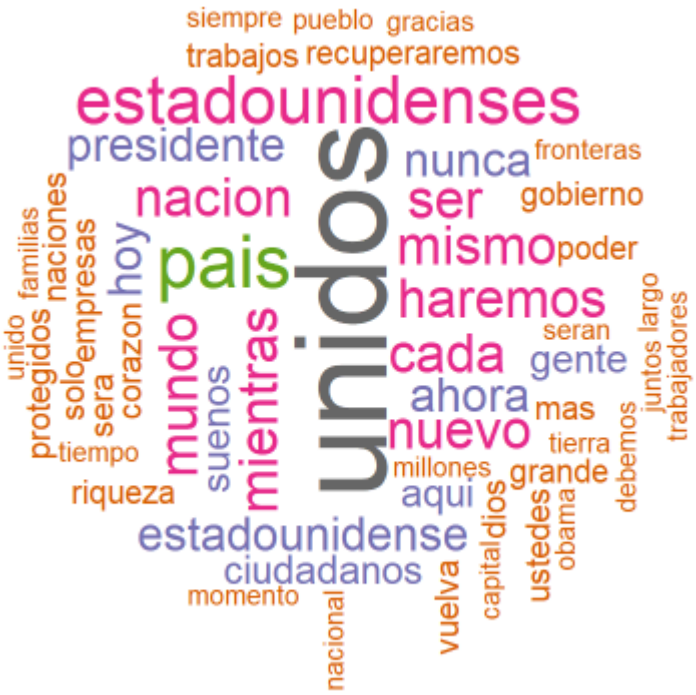
```
##                word freq
## freedom   freedom   13
## ring         ring   12
## dream       dream   11
## day           day   11
## let           let   11
## every       every    9
## one           one    8
## able         able    8
## together together    7
## nation     nation    4
```

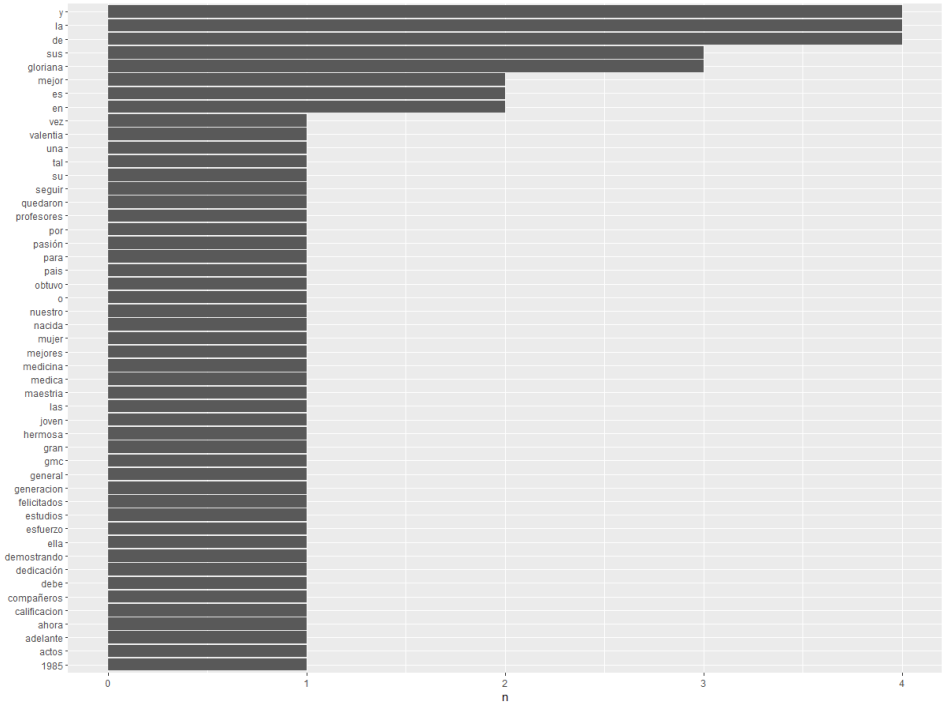
¿Cuál sería la diferencia entre un Corpus y una matriz de términos?

Índice

Análisis exploratorio

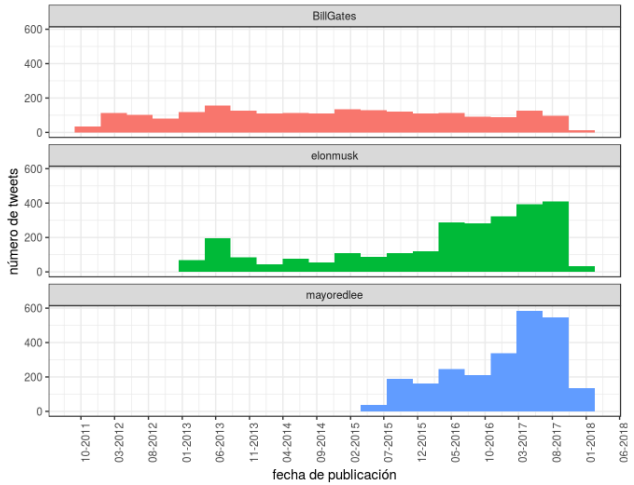
Para el análisis exploratorio, prácticamente se llevan a cado dos o tres: se cuentan la palabras, y se realizan nubes de palabras, y se pueden ver la asociación de acuerdo a ciertas palabras.

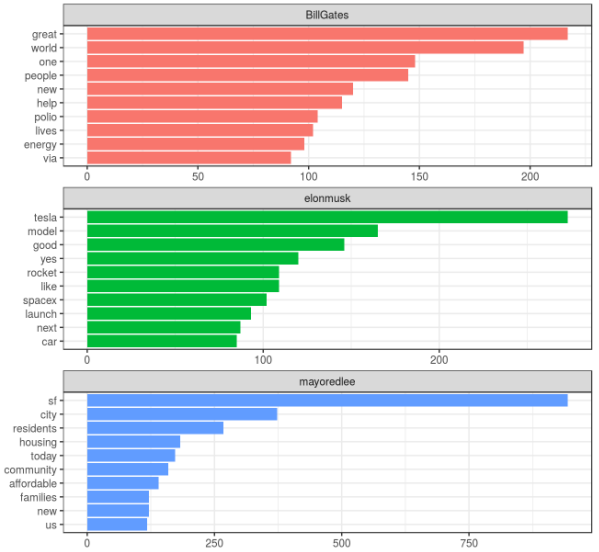


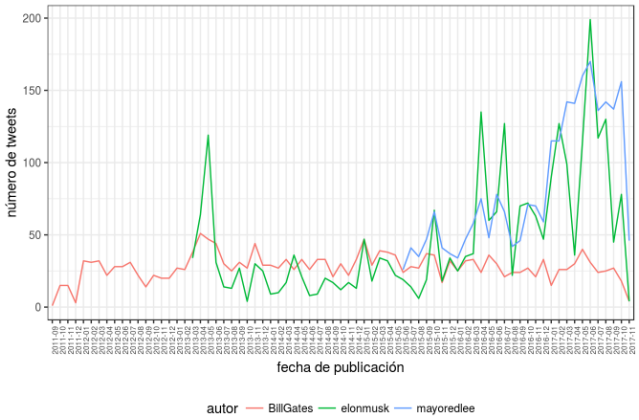


Análisis exploratorio

- Hay otros análisis descriptivos que se pueden hacer, si por ejemplo se obtiene información de Tweeter, dado que esta posee la fecha.







Índice

Análisis de sentimiento

Cómo bien se da a entender, el análisis de sentimiento trata de analizar los sentimientos o las características en dos partes: positivos y negativos.

Un análisis de sentimiento responde a:

¿Cuáles palabras han influido para determinar los sentimientos?

¿Qué sentimientos han sido predominantes? ¿Positivos, negativos? ¿Cómo han cambiado los sentimientos a través del tiempo?

El análisis de sentimiento, del trabajo de limpieza y tokenización, debe además estar basado en un lista llamada “Sentiment Lexicon”, que trata de catalogar a cada palabro según un estado, y así darle un puntuaje positivo o negativo.

Análisis de sentimiento

En nuestro caso, el Sentiment Lexicon usaremos el léxico Afinn. Este es un conjunto de palabras, puntuadas de acuerdo a qué tan positivamente o negativamente son percibidas. Las palabras que son percibidas de manera positiva tienen puntuaciones de -4 a -1; y las positivas de 1 a 4.

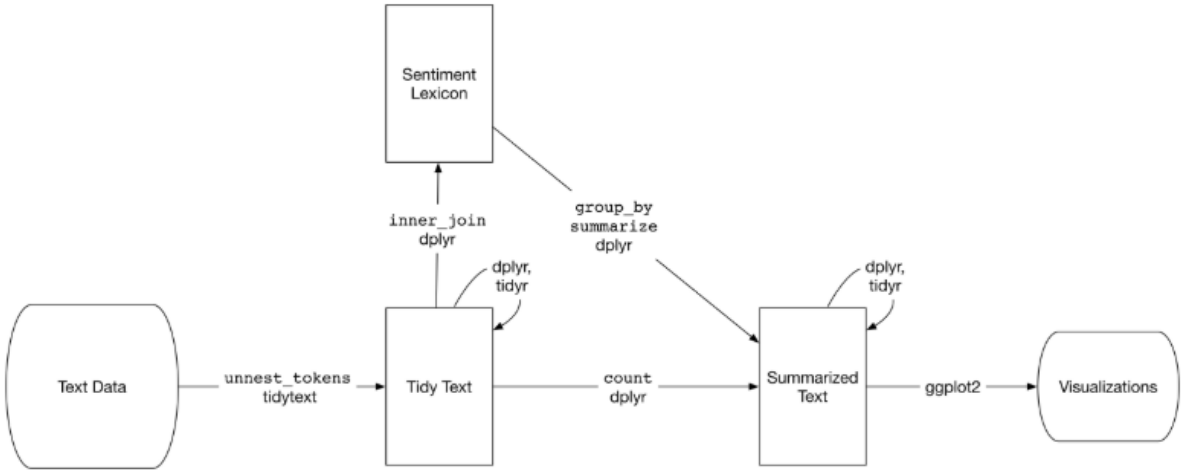
La versión que usaremos es una traducción automática, de inglés a español, de la versión del léxico presente en el conjunto de datos sentiments de *tidytext*, con algunas correcciones manuales. Por supuesto, esto quiere decir que este léxico tendrá algunos defectos, pero será suficiente para nuestro análisis. Descargamos este léxico de la siguiente dirección:

https://raw.githubusercontent.com/jboscomendoza/rpubs/master/sentimientos_afinn/lexico_afinn.en.es.csv

Análisis de sentimiento

```
## # A tibble: 2,476 x 3
##    Palabra     Puntuacion Word
##    <chr>            <int> <chr>
##  1 a bordo              1 aboard
##  2 abandona            -2 abandons
##  3 abandonado          -2 abandoned
##  4 abandonar           -2 abandon
##  5 abatido             -2 dejected
##  6 abatido             -3 despondent
##  7 aborrece            -3 abhors
##  8 aborrecer           -3 abhor
##  9 aborrecible         -3 abhorrent
## 10 aborrecido          -3 abhorred
## # ... with 2,466 more rows
```

Tenemos tres columnas. Una con palabras en español, su puntuación y una tercera columna con la misma palabra, en inglés.

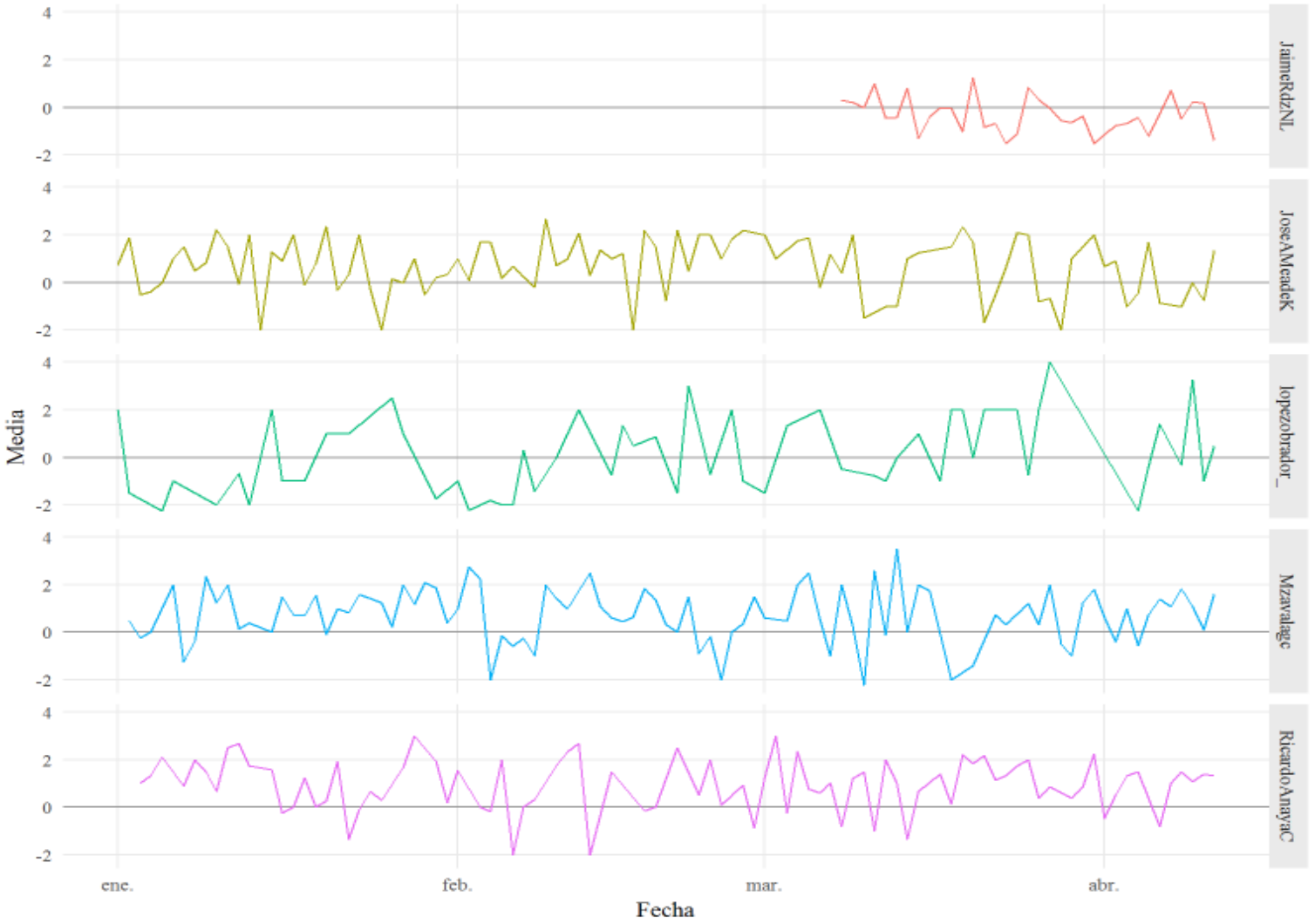


Análisis de sentimiento

El análisis de sentimiento puede analizar, según una persona, las palabras positivas y negativas.

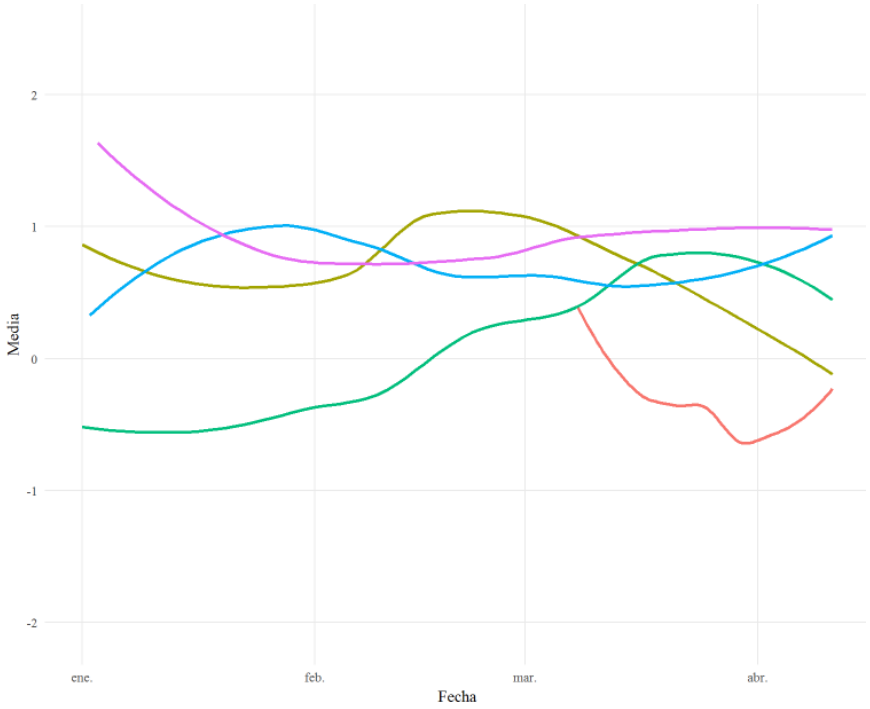
Análisis de sentimiento

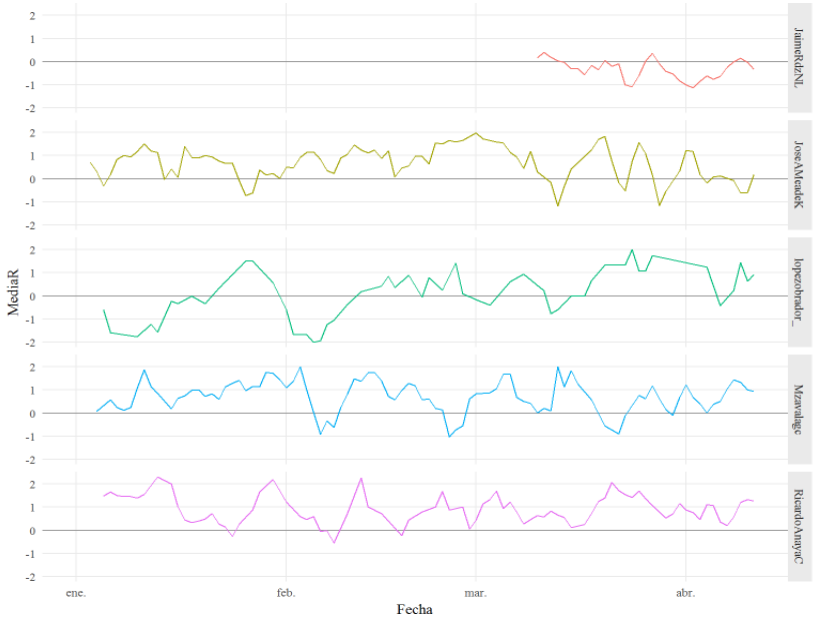
Se puede analizar los sentimientos, mediante puntuaciones, en el tiempo.



Análisis de sentimiento

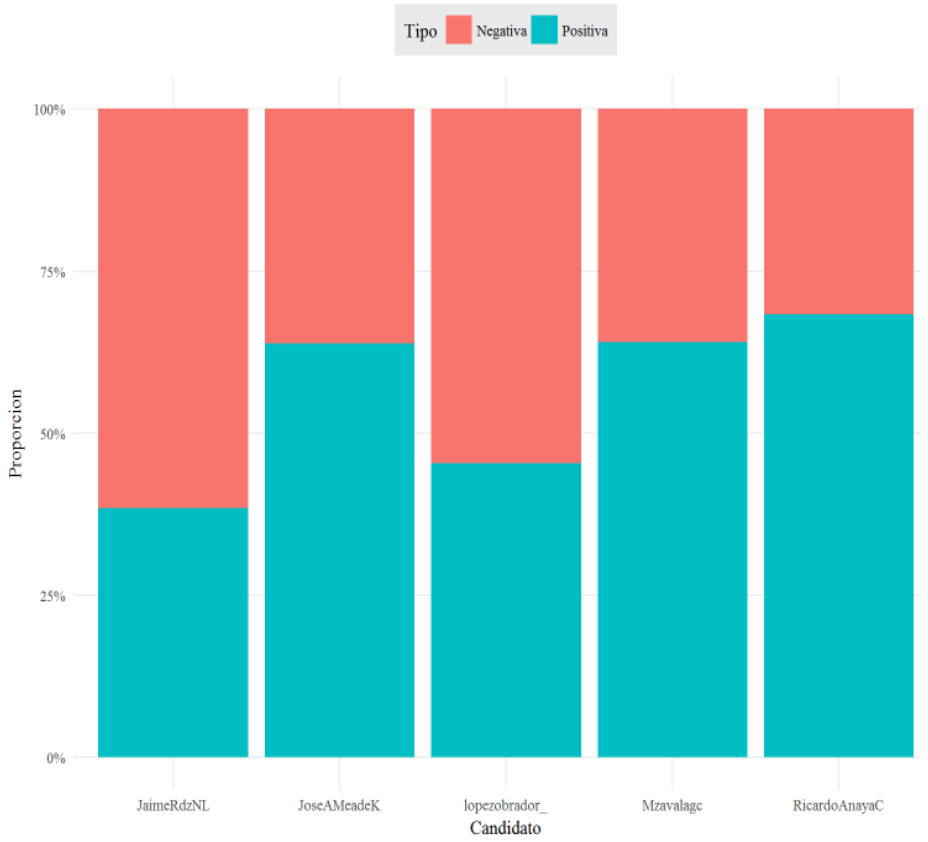
Si el interés fuera aproximar o describir dichos sentimientos, se podrían aplicar regresiones LOESS o métodos de medias móviles.





Análisis de sentimiento

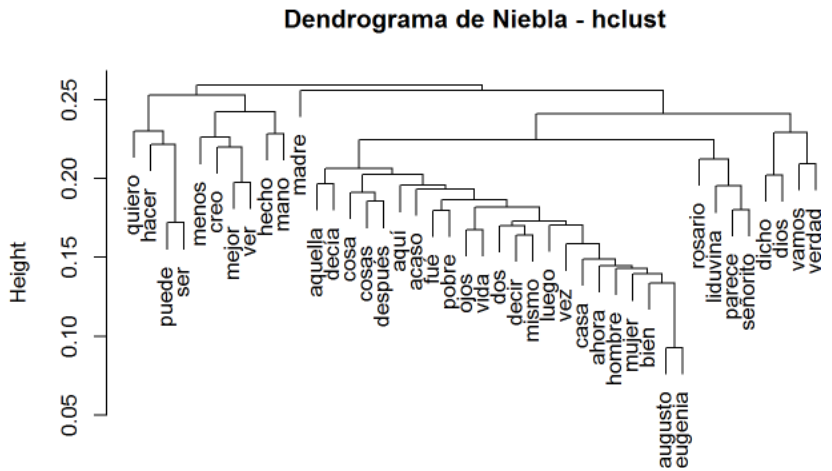
Se podrían comprar los sentimientos positivos y negativos, así como en el tiempo.

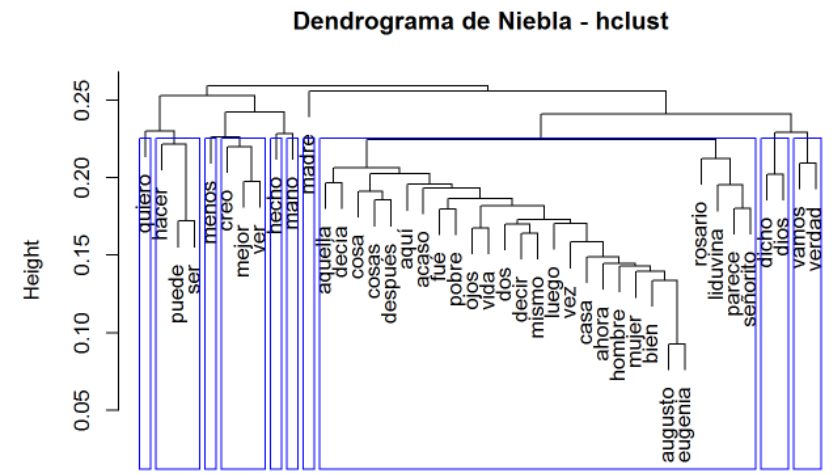


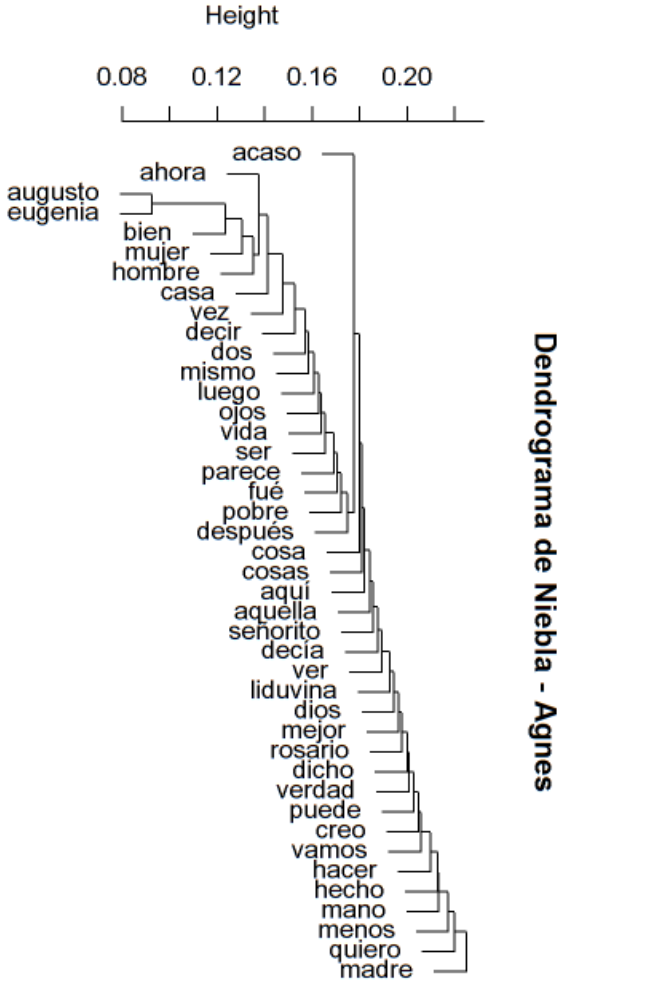
Índice

Otros análisis de TM

Podemos analizar la relación de las distancia entre las palabras. Para esto podemos realizar tanto agrupamientos por métodos jerárquicos como de k-medias.

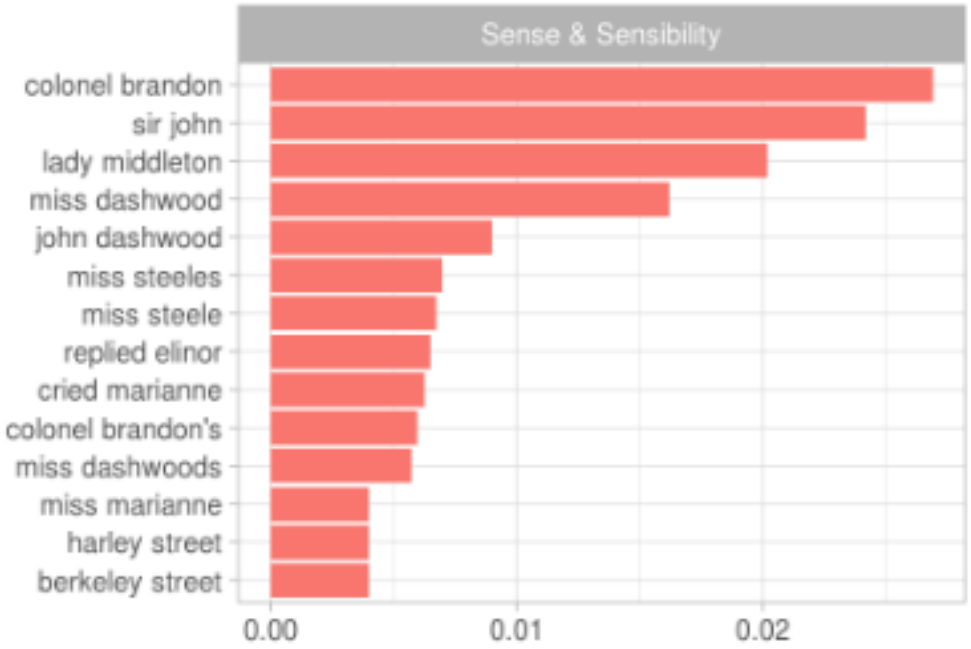


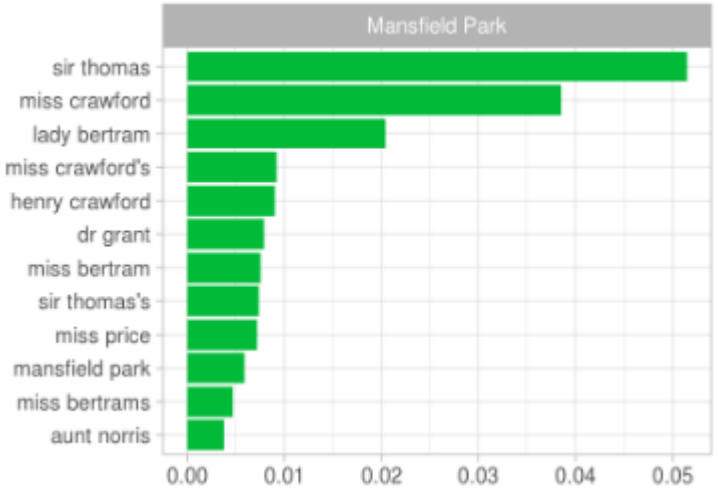


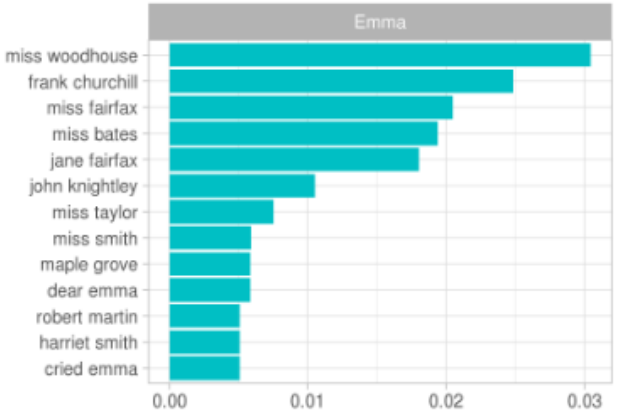


Otros análisis de TM

En vez de analizar una sola palabra, se podrían analizar dos palabras conjuntamente (o más), esto tanto para el análisis descriptivo, así como para el análisis de sentimiento. A esto lo denominamos como n-gramas. Lo más común es hacer un bi-grama.

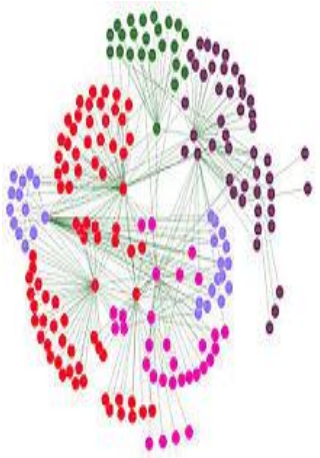




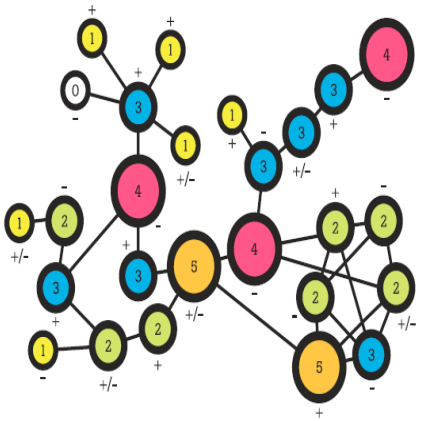


Otros análisis de TM

Es posible que estemos interesados en visualizar todas las relaciones entre las palabras simultáneamente, en lugar de solo las mejores a la vez. Como una visualización común, podemos organizar las palabras en una red o "gráfico". Aquí nos referiremos a un "gráfico" no en el sentido de una visualización, sino como una combinación de nodos conectados. Un gráfico se puede construir a partir de un objeto ordenado ya que tiene tres variables:



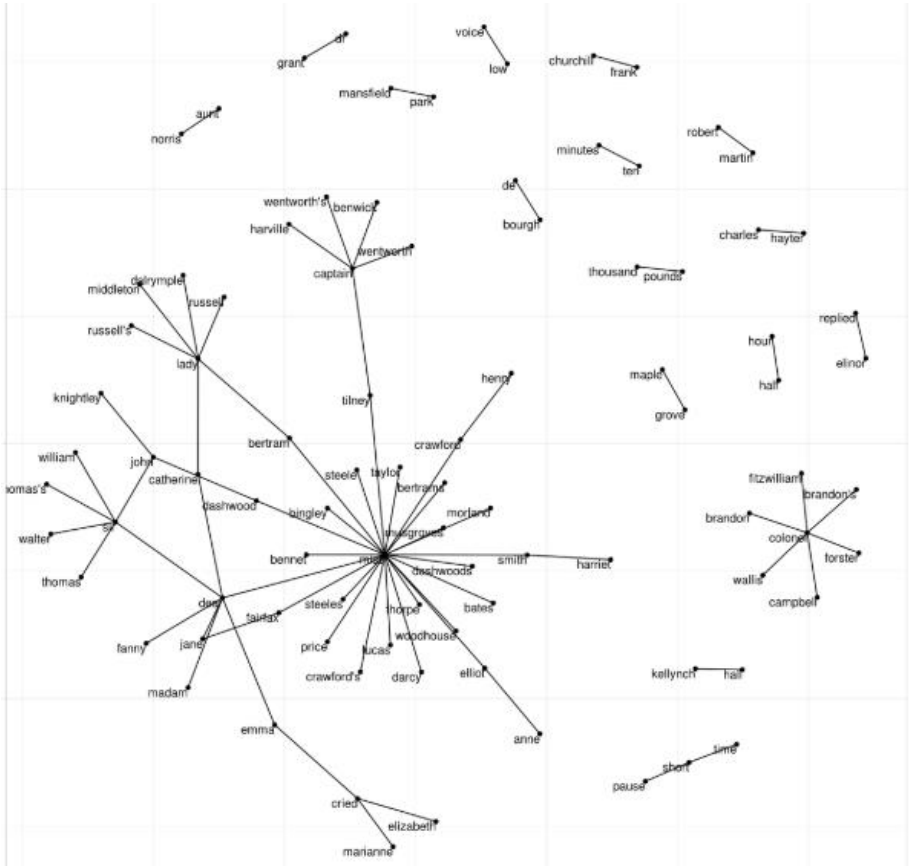
- **FROM**: el nodo del que viene una arista
- **TO**: el nodo hacia donde va un borde
- **WEIGHT**: un valor numérico asociado con cada borde

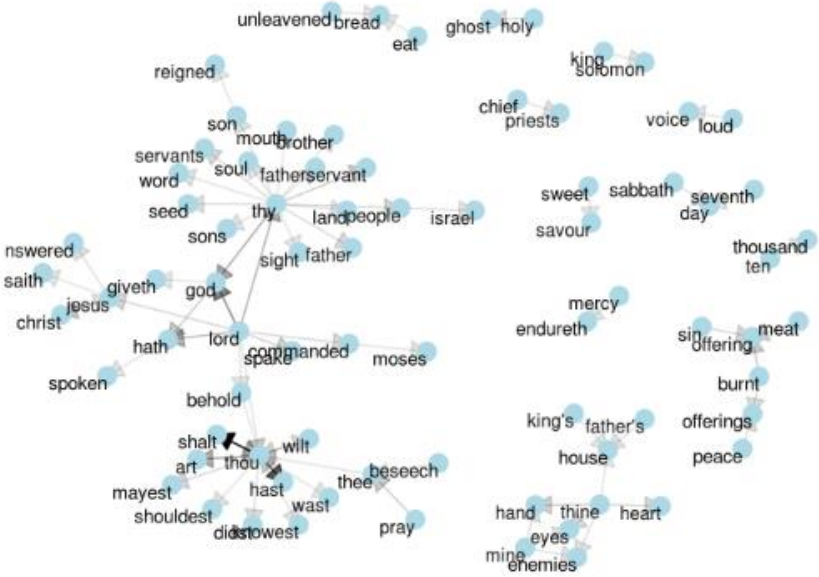


Realmente, aunque en el curso no hemos pasado por el análisis de las redes sociales para conocer las relaciones entre inviduos (acá palabras), este análisis rebela posibles formas de escritura y otra visión más profunda si se liga con el análisis de sentimiento.

Otros análisis de TM

Acá un ejemplo de redes sociales aplicado a palabras.





Índice

WEB SCRAPING

El Web Scraping

Web scraping es una técnica utilizada mediante programa de software para extraer información de sitios web. Usualmente, estos programas simulan la navegación de un humano en la WWW (World Wide Web) ya sea utilizando el protocolo HTPP manualmente, o incrustando un navegador en una aplicación.

El *web scraping* está muy relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda, Sin embargo, el *web scraping* se enfoca más en la transformación de datos sin estructura en la web en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento

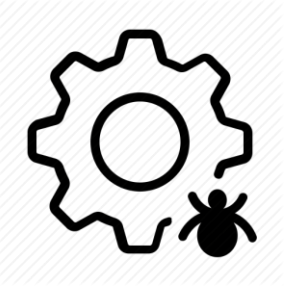
El Web Scraping

Mediante el Web Scraping podemos:

a. Obtener agregadores de contenido (colocar información en un solo sitio).
b. Reputación online mediante el análisis de sentimiento
c. Caza de tendencias (cool hunting), de lo que podría suceder en el futuro.
d. Optimización y estandarización de precios.
e. Monitorización de la competencia.
f. Optimización de ecommerce.
g. Etc, etc, etc.

El siguiente enlace explica de forma integral el Web Scraping:

https://aukera.es/blog/web-scraping/



Etapas de un análisis de Text Mining

A grandes rasgos, las etapas son:

1. Obtener los datos
2. Limpiar los datos
3. Tokenizar los datos (ver el Corpus)
4. Meter en un corpus y matriz de términos
5. Realizar estadísticas descriptivas
6. Análisis de sentimiento (Sentiment Lexicon)
7. Otros análisis
8. Concluir sobre el análisis

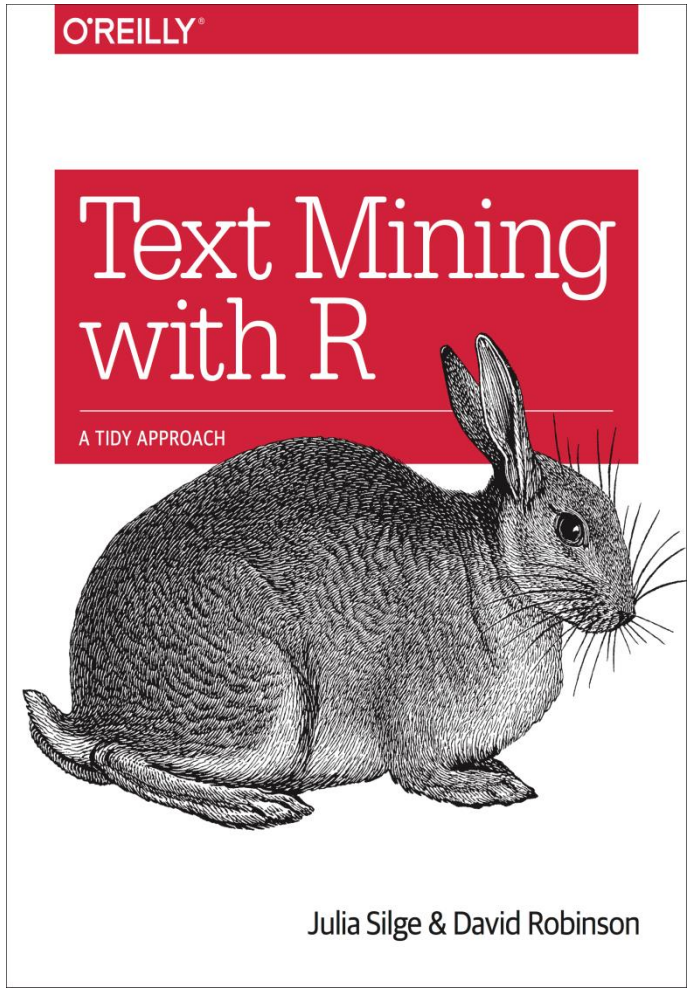
Conclusión: consulta de enlaces

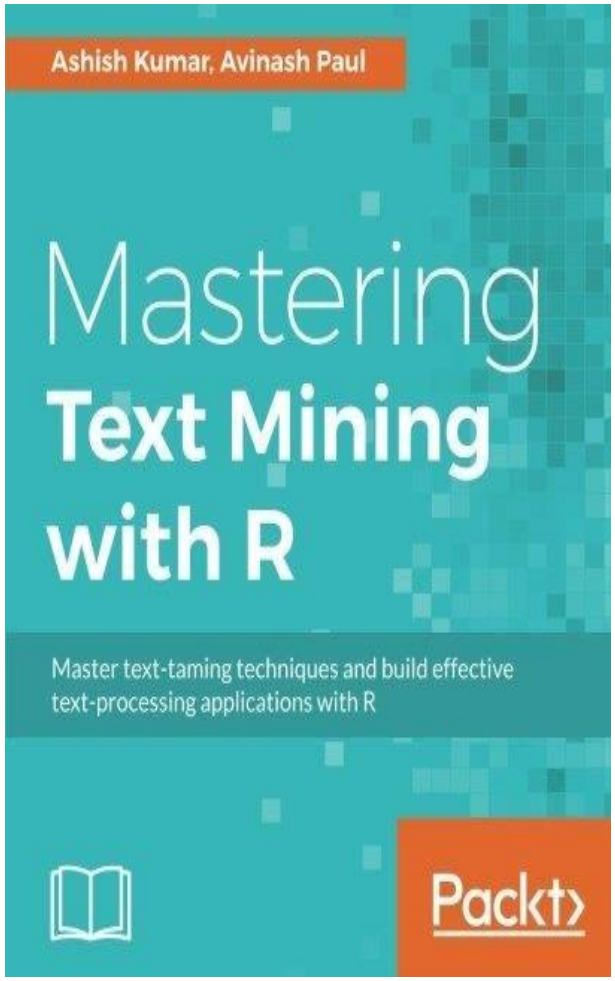
Para ampliar el tempa del Cluster Analysis, pueden consultar los siguientes enlaces:

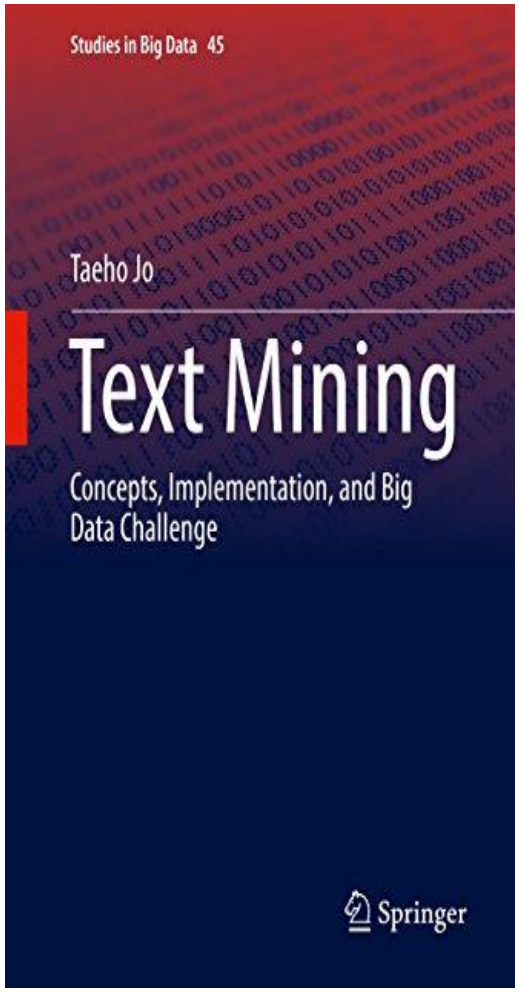
https://rpubs.com/Joaquin_AR/310338
https://rpubs.com/rdelgado/399475
http://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf
https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
https://uc-r.github.io/kmeans_clustering
https://www.kaggle.com/hendraherviawan/customer-segmentation-using-rfm-analysis-r
https://rpubs.com/vermaph/395036
https://www.r-bloggers.com/customer-segmentation-part-1-k-means-clustering/
https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html
https://towardsdatascience.com/how-to-cluster-your-customer-data-with-r-code-examples-6c7e4aa6c5b1
https://analyzecore.com/2015/02/16/customer-segmentation-lifecycle-grids-with-r/
http://www.rpubs.com/swapnilkura/ClusterAnalysis
https://rpubs.com/williamsurles/310847
https://rpubs.com/Damilolah/clustering
https://rpubs.com/shirokaner/320218
http://www.sthda.com/english/wiki/print.php?id=234
http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html

Conclusión: consulta de libros

Se recomienda la siguiente bibliografía:







Conclusión

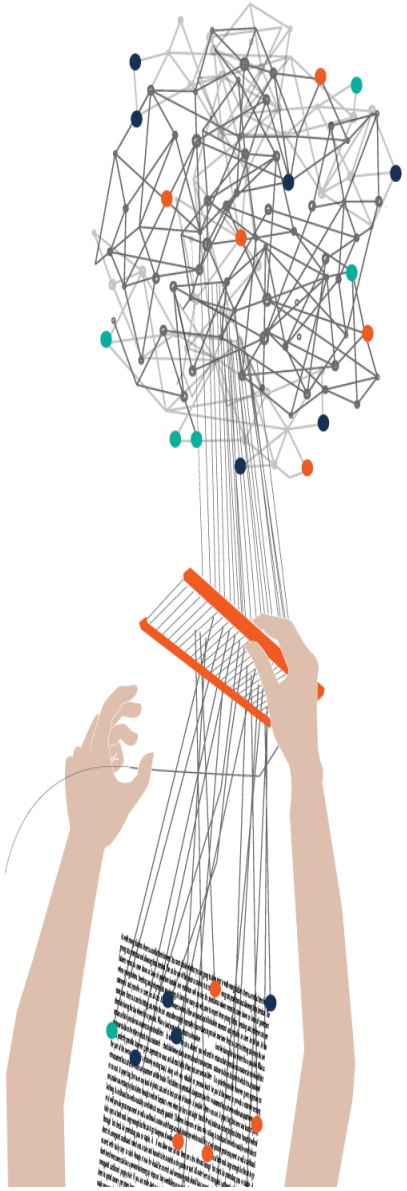
- El presente capítulo explica de forma breve el uso de la Minería de Texto como una forma de analizar las palabras.

- Se presentaron técnicas de lectura de datos, limpieza y tokenización, análisis descriptino, análisis de sentimiento y otras técnicas analíticas.

- En la actualidad la mayor potencia del Text Mining está en su uso con del Web Scraping.

- ¿Interezados en hacer Web Scraping con R?

https://www.datacamp.com/community/tutorials/r-web-scraping-rvest




The End