Tratamiento Inteligente de datos

(TID)

Prácticas de la asignatura

2018-2019

En colaboración con:

Participantes

Alejandro Campoy Nieves: alejandroac79@correo.ugr.es

Gema Correa Fernández: gecorrea@correo.ugr.es

Luis Gallego Quero: lgaq94@correo.ugr.es

Jonathan Martín Valera: jmv742@correo.ugr.es

Andrea Morales Garzón: andreamgmg@correo.ugr.es

Índice

Índice de figuras

Índice de cuadros

Descripción de los paquetes necesarios

A continuación, se describen los paquetes necesarios para el desarollo del proyecto:

- `tm` : Paquete específico para minería de datos, permite procesar datos de tipo texto. Se puede instalar usando : *install.packages("tm")*.
- `SnowballC` : Paquete adicional para minería de datos, implementa un algoritmo que permite reducir el número de términos con lo que trabajar, es decir, agrupa aquellos términos que contienen la misma raíz. El paquete soporta los siguientes idiomas: alemán, danés, español, finlandés, francés, húngaro, inglés, italiano, noruego, portugués, rumano, ruso, sueco y turco. Se puede instalar usando : *install.packages("SnowballC")*.
- `wordcloud` : Paquete para crear gráficas de nubes de palabras, permitiendo visualizar las diferencias y similitudes entre documentos. Se puede instalar usando : *install.packages("wordcloud")*.
- `arules` : Paquete que proporciona la infraestructura para representar, manipular y analizar datos y patrones de transacción (conjuntos de elementos frecuentes y reglas de asociación). Se puede instalar usando : *install.packages("arules")*.
- `arulesViz` : Paquete que extiende el paquete 'arules' con varias técnicas de visualización para reglas de asociación y conjuntos de elementos. El paquete también incluye varias visualizaciones interactivas para la exploración de reglas. Se puede instalar usando : *install.packages("arulesViz")*.
- `devtools` : Paquete que contiene una colección de herramientas de desarrollo de paquetes, usando conjuntamento con `rword2vec`, para obtener la agrupación de sinónimos. Se puede instalar usando : *install.packages("devtools")*.
- `rword2vec` : Paquete que toma un corpus de texto como entrada y produce los vectores de palabra como salida, usado especialmente para obtener las distancias que existen entre un término y los términos semejantes en el texto de formación (aprende la representación vectorial de las palabras). Se puede instalar usando : *install_github("mukul13/rword2vec")*.

1. Comprender el problema a resolver

El *dataset* **Drug Review Dataset**, proporcionado por *UCI Machine Learning Repository*, contiene una exhaustiva base de datos de medicamentos específicos, en la cual, el conjunto de datos muestra revisiones de pacientes sobre medicamentos específicos para unas condiciones particulares. Dichas revisiones se encuentran desglosadas en función del tema que se esté tratando: beneficios, efectos secundarios y comentarios generales. De igual modo, se dispone de una calificación de satisfacción general, es decir, de una calificación en base a los efectos secundarios del medicamento y de otra en base a la efectividad del mismo.

En este proyecto nos centraremos en el **análisis y experiencia qué tienen los usuarios con ciertos tipos de medicamentos**, para la realización y aplicación de las técnicas explicadas a lo largo del curso. Para ello, se proponen los siguientes objetivos principales:

- Realizar un análisis de sentimientos a partir de la experiencia de dichos usuarios en el uso de ciertos medicamentos, como por ejemplo ver la efectividad del medicamento cuánto está relacionado con los efectos secundarios o beneficios del mismo.
- Compatibilizar dicho modelo de datos con otros conjuntos de datos aportados en **Drugs.com**.

Las características de este conjunto de datos vienen descritas en la siguiente tabla 1:

Características del Data Set	Multivariable, texto
Características de los atributos	Entero
Tareas asociadas	Clasificación, regresión, clustering
Número de instancias	4143
Número de atributos	8
Valores vacíos	N/A
Área	N/A
Fecha de donación	10/02/2018
Veces visualizado	11047

Cuadro 1: Información del conjunto de datos

Los datos se dividen en un conjunto train (75 %) y otro conjunto test (25 %) y se almacenan en dos archivos *.tsv* (tab-separated-values), respectivamente. Los atributos que tenemos en este dataset son:

1. **urlDrugName** (categorical): nombre del medicamento/fármaco
2. **rating** (numerical): clasificación o puntuación del 1 a 10 del medicamento según el paciente
3. **effectiveness** (categorical): clasificación de la efectividad del medicamento según el paciente (5 posibles valores)
4. **sideEffects** (categorical): clasificación de los efectos secundarios del medicamento según el paciente (5 posibles valores)
5. **condition** (categorical): nombre de la condición (diagnóstico)
6. **benefitsReview** (text): opinión del paciente sobre los beneficios
7. **sideEffectsReview** (text): opinión del paciente sobre los efectos secundarios
8. **commentsReview** (text): comentario general del paciente

2. Preprocesamiento de datos

En este apartado, pondremos los datos a punto para la aplicación de diversas técnicas. Por tanto, para poder analizar dicho *dataset* y realizar el preprocesamiento al mismo, lo primero que se va hacer es leer el conjunto de datos *train* y *test*.

2.1. Lectura de datos

A continuación, mediante la función `read.table` procedemos a la lectura de los datos:

2.1.1. Lectura de datos train

Se va a procedeer a la lectura del conjunto de datos de entrenamiento.

```
# Lectura de datos train
datos_train <- read.table("datos/drugLibTrain_raw.tsv", sep="\t", comment.char="",
                          quote = "\"", header=TRUE)
```

Disponemos una matriz de 3107 filas x 9 columnas, asimismo vamos a ver un ejemplo de como distribuida la información, en donde para la fila tercera encontramos la siguiente información:

X	urlDrugName	rating	effectiveness	sideEffects	condition
1146	ponstel	10	Highly Effective	No Side Effects	menstrual cramps

Cuadro 2: Información contenida en una fila del conjunto de entrenamiento I

benefitsReview	sideEffectsReview	commentsReview
I was used to having cramps so badly that they would leave me balled up in bed for at least 2 days. The Ponstel doesn't take the pain away completely, but takes the edge off so much that normal activities were possible. Definitely a miracle medication!!	Heavier bleeding and clotting than normal.	I took 2 pills at the onset of my menstrual cramps and then every 8-12 hours took 1 pill as needed for about 3-4 days until cramps were over. If cramps are bad, make sure to take every 8 hours on the dot because the medication stops working suddenly and unfortunately takes about an hour to an hour and a half to kick back in.. if cramps are only moderate, taking every 12 hours is okay.

Cuadro 3: Información contenida en una fila del conjunto de entrenamiento II

De las tablas 2 y 3 podemos extraer que el medicamento **ponstel** con identificador **1146**, tiene la máxima puntuación por parte del paciente (**rating = 10**), el cual tiene un alto nivel de efectividad (**Highly Effective**) sin efectos secundarios (**No Side Effects**), usado para dolores menstruales (**menstrual cramps**), en donde el paciente dice que de estar tumbado en la cama con dolores ha pasado a poder realizar las actividades sin ningún impedimento. Además, asegura que tomar este medicamento le ha supuesto un sangrado más abundante y coagulación de lo normal. La dosis del medicamento oscila entre una píldora cada 8-12 horas durante 3-4 días.

2.1.2. Lectura de datos test

Se va a procedeer a la lectura del conjunto de datos de prueba.

```
# Lectura de datos test
datos_test <- read.table("./datos/drugLibTest_raw.tsv", sep="\t", comment.char="",
                         quote = "\"", header=TRUE)
```

Disponemos una matriz de 1036 filas x 9 columnas, asimismo vamos a ver un ejemplo de como distribuida la información, en donde para la fila primera encontramos la siguiente información:

De las tablas 4 y 5 podemos extraer que el medicamento **biaxin** con identificador **1366**, tiene una puntuación de 9 por parte del paciente (**rating = 9**), el cual tiene un nive considerable de efectividad (**Considerably Effective**) con efectos secundarios leves (**Mild Side Effects**), usado para la infección sinusal (**sinus**

infection), en donde el paciente dice que no está muy seguro de si el antibiótico ha destruido las bacterias que causan su infección sinusal. Además, asegura que tomar este medicamento le da algo de dolor de espalda y algunas náuseas. El paciente tomó los antibióticos durante 14 días y la infección sinusal desapareció al sexto día.

X	urlDrugName	rating	effectiveness	sideEffects	condition
1366	biaxin	9	Considerably Effective	Mild Side Effects	sinus infection

Cuadro 4: Información contenida en una fila del conjunto de prueba I

benefitsReview	sideEffectsReview	commentsReview
The antibiotic may have destroyed bacteria causing my sinus infection. But it may also have been caused by a virus, so its hard to say.	Some back pain, some nauseau.	Took the antibiotics for 14 days. Sinus infection was gone after the 6th day.

Cuadro 5: Información contenida en una fila del conjunto de prueba II

La representación del documento se llevará a cabo utilizando palabras, después de un debido filtrado para minimizar la dimensión del espacio de trabajo.

2.2. Procesamiento de los datos

Dado que la representación total del documento puede tener una alta dimensión, se va a procedeer a construir un corpus, necesario para la aplicación de métodos de limpieza y esructuración del texto de entrada e identificación de un subconjunto simplificado de las características del documento con el fin de poder ser representado en un análisis posterior.

2.2.1. Eliminar columnas

El primer paso que vamos a realizar es la **eliminación de columnas**, las cuales contienen información irrelevante para nuestro análisis.

Eliminar columna ID

Al conjunto de datos utilizado se le ha añadido de forma automática una novena columna, que representa un ID para cada uno de los datos con los que estamos trabajando. Como este ID no nos aporta información alguna, hemos decidido quitarla directamente del *dataframe*. Esta columna se corresponde con la primera columna, por lo cuál, debemos eliminar la columna que se corresponde con la posición 1. Los cambios que hacemos en el *dataset* deben modificarse tanto en el conjunto de test como el de train, para que los resultados sean consistentes.

```
datos_train = datos_train[-1] # Eliminar columna para el ID en el train
datos_test = datos_test[-1] # Eliminar columna para el ID en el test
```

Eliminar columna de commentsReview

Consideramos que la información contenida en *commentsReview* no es de nuestro interés. En este atributo se almacena texto, en el cual los consumidores de los medicamentos suelen poner en la mayoría de casos la frecuencia o la dosis con la que consumen la misma. En otros casos menos frecuentes, se establecen comentarios más arbitrarios en el que se muestran sus sensaciones o información sin relevancia. Incluso en algunos casos este campo aparece vacío. Es por eso, que hemos decidido eliminar la columna, tanto para el conjunto test como el train.

```
datos_train = datos_train[-8] # Eliminar columna para el commentsReview en el train
datos_test = datos_test[-8] # Eliminar columna para el commentsReview en el test
```

2.2.2. Categorización de variables

Para poder analizar y trabajar más fácilmente con la información de *sideEffects* y *effectiveness*, se va a realizar una conversión de dichas columnas a forma cuantitativa, es decir, vamos asignar una etiqueta numérica a cada valor pertinente, tanto para para *train* como *test*.

A continuación, vamos a cuantificar la columna de *sideEffects*, para ello se añade una nueva columna a nuestro conjunto de datos denominada *sideEffectsNumber* que nos clasifica los posibles valores de la columna *sideEffects* en un rango numérico, comprendido entre 1 y 5. Dicha columna hace referencia a la clasificación de los efectos secundarios del medicamento según el paciente, en donde la etiqueta con valor 1 hará referencia a que no haya ningún efecto secundario y la etiqueta con valor 5 a que tiene efectos secundarios extremadamente graves:

- Extremely Severe Side Effects (efectos secundarios extremadamente graves) : 5
- Severe Side Effects (efectos secundarios graves): 4
- Moderate Side Effects (efectos secundarios moderados) : 3
- Mild Side Effects (efectos secundarios leves) : 2
- No Side Effects (sin efectos secundarios) : 1

```
# Datos Train
datos_train$sideEffectsNumber[datos_train$sideEffects=="Extremely Severe Side Effects"]<-5
datos_train$sideEffectsNumber[datos_train$sideEffects=="Severe Side Effects"]<-4
datos_train$sideEffectsNumber[datos_train$sideEffects=="Moderate Side Effects"] <- 3
datos_train$sideEffectsNumber[datos_train$sideEffects=="Mild Side Effects"]<- 2
datos_train$sideEffectsNumber[datos_train$sideEffects=="No Side Effects"]<- 1

# Datos Test
datos_test$sideEffectsNumber[datos_test$sideEffects=="Extremely Severe Side Effects"]<-5
datos_test$sideEffectsNumber[datos_test$sideEffects=="Severe Side Effects"]<-4
datos_test$sideEffectsNumber[datos_test$sideEffects=="Moderate Side Effects"]<-3
datos_test$sideEffectsNumber[datos_test$sideEffects=="Mild Side Effects"]<-2
datos_test$sideEffectsNumber[datos_test$sideEffects=="No Side Effects"]<-1
```

Podemos comprobar que se ha creado la nueva columna *sideEffectsNumber*, y que se han añadido los cambios comentados anteriormente.

```
head(datos_train$sideEffectsNumber, 10)
```

```
##  [1] 2 4 1 2 4 4 2 1 1 5
```

Volvemos a aplicar el mismo procedimiento para la columna de *effectiveness*, creándonos para ello una columna denominada *effectivenessNumber*. Dicha columna, hace referencia a la clasificación de la efectividad del medicamento según el paciente, en donde la etiqueta con valor 1 hace referencia a que el medicamente es ineficaz y la etiqueta con valor 5 a que el medicamente es altamente eficaz:

- Highly Effective (altamente efectivo): 5
- Considerably Effective (considerablemente efectivo) : 4
- Moderately Effective (moderadamente efectivo) : 3
- Marginally Effective (marginalmente efectivo) : 2
- Ineffective (ineficaz) : 1

```
# Datos de entrenamiento
datos_train$effectivenessNumber[datos_train$effectiveness=="Highly Effective"]<-5
datos_train$effectivenessNumber[datos_train$effectiveness=="Considerably Effective"]<-4
```

```
datos_train$effectivenessNumber[datos_train$effectiveness=="Moderately Effective"]<-3
datos_train$effectivenessNumber[datos_train$effectiveness=="Marginally Effective"]<-2
datos_train$effectivenessNumber[datos_train$effectiveness=="Ineffective"]<- 1

# Datos de test
datos_test$effectivenessNumber[datos_test$effectiveness=="Highly Effective"]<-5
datos_test$effectivenessNumber[datos_test$effectiveness=="Considerably Effective"]<-4
datos_test$effectivenessNumber[datos_test$effectiveness=="Moderately Effective"]<-3
datos_test$effectivenessNumber[datos_test$effectiveness=="Marginally Effective"]<-2
datos_test$effectivenessNumber[datos_test$effectiveness=="Ineffective"]<-1
```

Comprobamos que se ha creado la nueva columna *effectivenessNumber*, y que se han añadido los nuevos cambios.

```
head(datos_train$effectivenessNumber, 10)
```

```
##  [1] 5 5 5 2 2 1 5 4 5 1
```

Por tanto, una vez eliminadas las columnas anteriores y modificadas las necesarias, ya podemos continuar con el procesamiento de los datos. Para ello, lo primero tenemos que hacer es cargar la librería que procesa los datos de tipo texto en R, para la construcción y manipulación del corpus. La librería más conocida se llama **tm**, aunque también haremos uso del paquete **SnowballC** para realizar el *Stemming*. Si no tenemos instaladas las librerías:

```
# Paquete para minería de datos, permite procesar datos de tipo texto
# El paquete tm, necesita el paquete NLP
library("NLP")
library("tm")

# Paquete para minería de datos, agrupa aquellos términos que contienen la misma raíz
library("SnowballC")
```

2.2.3. Creación del corpus

Para poder obtener la estructura con la que vamos a procesar nuestra información, debemos obtener un vector con documentos. En nuestro caso, cada uno de los documentos se corresponde con una opinión sobre un fármaco (*benefitsReview*) y los efectos que tiene (*sideEffectsReview*). Para ello, primero debemos de construir un vector con todas los opiniones del *dataframe* y convertir cada elemento del vector al formato de documento. Podemos usar la función *VectorSource* para hacer esta conversión. Se deberán realizar todas las modificaciones tanto para el conjunto train como test.

```
# Datos train

# Nos quedamos con la única columna del dataset que nos interesa.
# Necesitamos obtenerla en forma de vector, y no como un dataframe de una columna,
# por lo que usamos as.vector para hacer la conversión
benefits_train_review_data = as.vector(datos_train$benefitsReview)
effects_train_review_data = as.vector(datos_train$sideEffectsReview)

# Lo convertimos en la estructura de documento, y lo guardamos ya en el corpus
# que lo vamos a utilizar
benefits_train_corpus = (VectorSource(benefits_train_review_data))
effects_train_corpus = (VectorSource(effects_train_review_data))

# Creamos el propio corpus
```

```
benefits_train_corpus <- Corpus(benefits_train_corpus)
effects_train_corpus <- Corpus(effects_train_corpus)
```

```
# Datos test

# Nos quedamos con la única columna del dataset que nos interesa.
# Necesitamos obtenerla en forma de vector, y no como un dataframe de una columna,
# por lo que usamos as.vector para hacer la conversión
benefits_test_review_data = as.vector(datos_test$benefitsReview)
effects_test_review_data = as.vector(datos_test$sideEffectsReview)

# Lo convertimos en la estructura de documento, y lo guardamos ya en el corpus
# que lo vamos a utilizar
benefits_test_corpus = (VectorSource(benefits_test_review_data))
effects_test_corpus = (VectorSource(effects_test_review_data))

# Creamos el propio corpus
benefits_test_corpus <- Corpus(benefits_test_corpus)
effects_test_corpus <- Corpus(effects_test_corpus)
```

Podemos ver que funciona accediendo a uno cualquiera, de la forma `inspect(benefits_train_corpus[4])`:

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

[1] The acid reflux went away    months  just   days     drug. The heartburn started   soon  I stopped taking . So I began treatment . 6
months passed  I stopped taking . The heartburn came back,  seemed worse even. The doctor said I  try another 6 month treatment. I ,
exact thing happened. This went    three years. I asked    curing  reflux. The doctor quite frankly told      cure,   "treatment
symptoms". I  told  I  probably      rest   life.
```

Figura 1: Contenido de benefits_train_corpus

O de la forma `benefits_train_corpus[[4]]$content`:

```
[1] "The acid reflux went away    months  just   days     drug. The heartburn started   soon  I stopped taking . So I began treatment . 6
months passed  I stopped taking . The heartburn came back,  seemed worse even. The doctor said I  try another 6 month treatment. I ,
exact thing happened. This went    three years. I asked    curing  reflux. The doctor quite frankly told      cure,   \"treatment
symptoms\". I  told  I  probably      rest   life."
```

Figura 2: Contenido de benefits_train_corpus

Y si nos fijamos en el contenido, vemos que tiene signos de puntuación y exclamación.

2.2.4. Eliminar signos de puntuación

Como hemos podido ver en el documento que se ha mostrado por pantalla, en él se aprecia el uso de signos de puntuación y exclamación. En un principio, no tiene sentido en *Data Mining* contemplar los signos de puntuación, ya que no nos van a aportar información. Por ello, los quitamos, como se puede ver a continuación. Con `tm_map(corpus, removePunctuation)`, se eliminan los símbolos: ! " $ % & ‘() * + , - . / : ; < = > ? @ [ ] ^ _’ { | } ~

```
# Una vez que tenemos el corpus creado, continuamos con el procesamiento para los datos train
benefits_train_corpus <- tm_map(benefits_train_corpus, content_transformer(removePunctuation))
effects_train_corpus <- tm_map(effects_train_corpus, content_transformer(removePunctuation))

# Una vez que tenemos el corpus creado, continuamos con el procesamiento para los datos test
```

```
benefits_test_corpus <- tm_map(benefits_test_corpus, content_transformer(removePunctuation))
effects_test_corpus <- tm_map(effects_test_corpus, content_transformer(removePunctuation))
```

Si volvemos a mostrar la opinión número cuatro, vemos como todos los signos han desaparecido. De hecho, podemos inspeccionar el corpus, y se ve como todos los signos de puntuación, exclamación y derivados ya no están.

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

[1] The acid reflux went away for a few months after just a few days of being on the drug The heartburn started again as soon as I stopped
taking it So I began treatment again 6 months passed and I stopped taking it The heartburn came back and seemed worse even The doctor said
I should try another 6 month treatment I did and the same exact thing happened This went on for about three years I asked why this wasnt
curing my reflux The doctor quite frankly told me that it wasnt a cure but a treatment for the symptoms I was told that I would probably
be on it for the rest of my life
```

Figura 3: Contenido de benefits_train_corpus con inspect(benefits_train_corpus[4])

Ocurre lo mismo con el comentario de efectos número siete.

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

[1] a few experiences of nausiea heavy moodswings on the days I do not take it decreased appetite and some negative affect on my shortterm
memory
```

Figura 4: Contenido de effects_train_corpus con inspect(effects_train_corpus[7])

2.2.5. Conversión de las mayúsculas en minúsculas

Para poder hacer uso de los términos por igual, debemos convertir las mayúsculas en minúsculas. Ya que normalmente se convierte en minúsculas todas las letras para que los comienzos de oración no sean tratados de manera diferente por los algoritmos.

```
benefits_train_corpus <- tm_map(benefits_train_corpus, content_transformer(tolower))
#inspect(benefits_train_corpus[4])

effects_train_corpus <- tm_map(effects_train_corpus, content_transformer(tolower))
#inspect(effects_train_corpus[7])

benefits_test_corpus <- tm_map(benefits_test_corpus, content_transformer(tolower))
effects_test_corpus <- tm_map(effects_test_corpus, content_transformer(tolower))
```

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

[1] the acid reflux went away for a few months after just a few days of being on the drug the heartburn started again as soon as i stopped
taking it so i began treatment again 6 months passed and i stopped taking it the heartburn came back and seemed worse even the doctor said
i should try another 6 month treatment i did and the same exact thing happened this went on for about three years i asked why this wasnt
curing my reflux the doctor quite frankly told me that it wasnt a cure but a treatment for the symptoms i was told that i would probably
be on it for the rest of my life
```

Figura 5: inspect(benefits_train_corpus[4])

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

[1] a few experiences of nausiea heavy moodswings on the days i do not take it decreased appetite and some negative affect on my shortterm
memory
```

Figura 6: inspect(effects_train_corpus[7])

2.2.6. Eliminación de Stopwords

En cualquier idioma, hay palabras que son tan comunes o muy utilizadas que no aportan información relevante, a dichas palabras se las conoce como *stopwords* o palabras *stop*. Por ejemplo, en español, las palabras "la", "a", "en", "de" son ejemplos de *stopwords*. Este tipo de palabras debemos de suprimirlas de nuestro corpus. Como, en nuestro caso, el contenido del corpus está en inglés, debemos especificar el idioma correcto para que nos elimine del corpus las palabras adecuadas en dicho idioma.

```
benefits_train_corpus <- tm_map(benefits_train_corpus, content_transformer(removeWords), stopwords("eng
#inspect(benefits_train_corpus[4])

effects_train_corpus <- tm_map(effects_train_corpus, content_transformer(removeWords), stopwords("englis
inspect(effects_train_corpus[7])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
## [1]   experiences  nausiea heavy moodswings   days    take  decreased appetite   negative affect   sh
```

```
benefits_test_corpus <- tm_map(benefits_test_corpus, content_transformer(removeWords), stopwords("englis
effects_test_corpus <- tm_map(effects_test_corpus, content_transformer(removeWords), stopwords("english
```

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

[1]  acid reflux went away    months  just   days     drug  heartburn started   soon   stopped taking    began treatment  6 months passed
stopped taking   heartburn came back  seemed worse even  doctor said   try another 6 month treatment      exact thing happened  went
three years  asked   wasnt curing  reflux  doctor quite frankly told    wasnt  cure   treatment   symptoms   told    probably      rest
life
```

Figura 7: inspect(benefits_train_corpus[4])

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

[1]   experiences  nausiea heavy moodswings   days    take  decreased appetite   negative affect   shortterm memory
```

Figura 8: inspect(effects_train_corpus[7])

Ahora ya hemos eliminado las stopwords de forma correcta.

2.2.7. Agrupación de sinónimos:

Con el fin de disminuir la dimensión del espacio a trabajar, se pueden identificar palabras distintas con el mismo significado y reemplazarlas por una sola palabra. Para ello se toman los sinónimos de dicha palabra. Dentro de las librerías que podemos usar para agrupar sinónimos, destacamos dos: `wordnet` y `rword2vec`. Sin embargo, por su sencillez se va hacer uso de `rword2vec`. Previamente, se obtendrán que palabras son las

que mayor frecuencia presentan en nuestro texto, para ello nos quedamos con las 100 más representativas tanto para *benefitsReview* como *sideEffectsReview* del conjunto train y test:

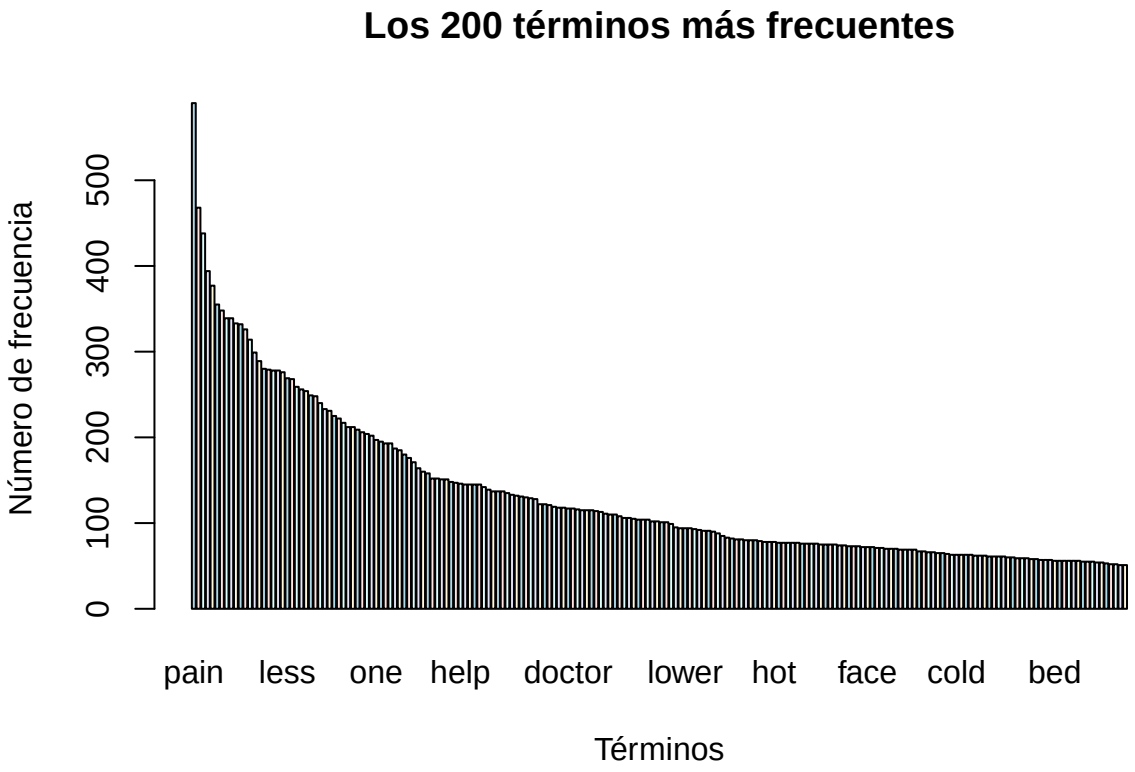
```
# Columna benefitsReview del conjunto train

# Obtenemos su matriz de términos
matrix_train_benefits_corpus <- TermDocumentMatrix(benefits_train_corpus)
# No tenemos los datos en la matriz que buscamos, sino en un vector
# por tanto, lo convertimos en matriz
matrix_train_benefits_corpus <- as.matrix(matrix_train_benefits_corpus)
# Sumamos las filas para obtener la frecuencia de una palabra en benefitsReview
matrix_train_benefits_corpus <- rowSums(matrix_train_benefits_corpus)
# Ordenamos de mayor a menor los términos y nos quedamos con lso 100 primeros
terms_frecuency_benefits_train_corpus <- sort(matrix_train_benefits_corpus, decreasing = TRUE)
terms_frecuency_benefits_train_corpus_200 <- terms_frecuency_benefits_train_corpus[1:200]
terms_frecuency_benefits_train_corpus_200
```

```
##          pain        taking          drug          also           day
##           590           468           438           394           377
##          skin         sleep     treatment          able    medication
##           355           348           339           339           333
##          time        better       effects          take          much
##           332           326           314           299           289
##          feel           get      symptoms    depression       helped
##           280           279           278           278           276
##          less          side         years       reduced      benefits
##           269           268           259           256           254
##          days          acne         first          felt       without
##           249           248           240           233           231
##       anxiety          life        within         still          like
##           225           222           217           212           212
##           now     effective          took       started           one
##           209           206           204           202           197
##          work          back        months         night         weeks
##           195           193           193           187           185
##         blood          well           can       stopped        severe
##           180           176           171           164           160
##          just          used          even        normal           two
##           158           152           152           151           151
##         taken       feeling          help      improved          mood
##           148           147           146           145           145
##         hours     increased          made      pressure       control
##           145           145           142           139           137
##          good        didnt          went          hair       cleared
##           137           137           135           133           132
##        energy          week        really           use    completely
##           131           130           129           128           122
##         daily        almost        weight    prescribed     reduction
##           122           121           119           118           118
##        doctor     infection          dose          away     decreased
##           117           117           116           115           115
##         times           due        relief          none          long
##           115           114           113           111           110
##       however          year          will         since         every
##           110           108           106           106           105
```

```
##        effect          dont        little        longer       patient
##           104           104           104           102           102
##          loss         never        worked       noticed        seemed
##           101           101            99            95            94
##         lower       benefit         using   experienced         month
##            94            94            93            92            91
##     headaches         tried         think          gone           got
##            91            90            88            85            83
##          high        period      medicine       overall          pill
##            82            81            81            80            80
##           ive       getting         great          stop           hot
##            80            79            78            78            78
##         began          many      migraine          mild   improvement
##            77            77            77            77            77
##        reduce      decrease        things        levels        dosage
##            76            76            76            76            75
##      problems         found         going       results          need
##            75            75            75            74            74
##          acid        became         focus    experience          face
##            73            73            73            72            72
##         clear        around     condition         panic      function
##            72            71            71            70            70
##       several         helps         level       morning   cholesterol
##            70            69            69            69            69
##         heart       quickly        asleep significantly        though
##            67            67            66            66            65
##          make       attacks          know          cold           per
##            65            64            63            63            63
##       problem           see      although      increase    eliminated
##            63            63            62            62            62
##         three         works          eyes           bad   immediately
##            61            61            61            61            60
##      headache       usually    difference        caused         worse
##            60            59            59            59            58
##       flashes      starting       stomach       ability           bed
##            58            57            57            57            56
##        reflux     migraines         drugs           dry       allowed
##            56            56            56            56            56
##       another           new          gave          body          easy
##            55            55            55            54            54
##       smoking          last       lowered        enough           etc
##            53            52            52            51            51
```

Y visualizamos dichos términos gráficamente:

```
graph_terms_frecuency_benefits_train_corpus <- as.matrix(terms_frecuency_benefits_train_corpus_200)
barplot(graph_terms_frecuency_benefits_train_corpus[1:200,],  xlab="Términos", ylab="Número de frecuenc
        col = c("lightblue", "mistyrose", "lightcyan", "lavender", "cornsilk"))
title(main = list("Los 200 términos más frecuentes", font = 2))
```



Una vez que tenemos los 100 términos con mayor frecuencia en nuestra columna *benefitsReview* y su frecuencia asociada, pasamos a matriz dichos datos, con el fin de obtener solo las palabras y descartar su frecuencia.

```
# Convertimos a matriz "terms_frecuency_benefits_corpus_100"
terms_frecuency_benefits_train_corpus_200 <- as.matrix(terms_frecuency_benefits_train_corpus_200)
terms_frecuency_benefits_train_corpus_200
```

```
##                [,1]
## pain            590
## taking          468
## drug            438
## also            394
## day             377
## skin            355
## sleep           348
## treatment       339
## able            339
## medication      333
## time            332
## better          326
## effects         314
## take            299
## much            289
## feel            280
## get             279
## symptoms        278
## depression      278
## helped          276
## less            269
## side            268
```

```
## years          259
## reduced        256
## benefits       254
## days           249
## acne           248
## first          240
## felt           233
## without        231
## anxiety        225
## life           222
## within         217
## still          212
## like           212
## now            209
## effective      206
## took           204
## started        202
## one            197
## work           195
## back           193
## months         193
## night          187
## weeks          185
## blood          180
## well           176
## can            171
## stopped        164
## severe         160
## just           158
## used           152
## even           152
## normal         151
## two            151
## taken          148
## feeling        147
## help           146
## improved       145
## mood           145
## hours          145
## increased      145
## made           142
## pressure       139
## control        137
## good           137
## didnt          137
## went           135
## hair           133
## cleared        132
## energy         131
## week           130
## really         129
## use            128
## completely     122
## daily          122
```

```
## almost         121
## weight         119
## prescribed     118
## reduction      118
## doctor         117
## infection      117
## dose           116
## away           115
## decreased      115
## times          115
## due            114
## relief         113
## none           111
## long           110
## however        110
## year           108
## will           106
## since          106
## every          105
## effect         104
## dont           104
## little         104
## longer         102
## patient        102
## loss           101
## never          101
## worked          99
## noticed         95
## seemed          94
## lower           94
## benefit         94
## using           93
## experienced     92
## month           91
## headaches       91
## tried           90
## think           88
## gone            85
## got             83
## high            82
## period          81
## medicine        81
## overall         80
## pill            80
## ive             80
## getting         79
## great           78
## stop            78
## hot             78
## began           77
## many            77
## migraine        77
## mild            77
## improvement     77
```

```
## reduce          76
## decrease        76
## things          76
## levels          76
## dosage          75
## problems        75
## found           75
## going           75
## results         74
## need            74
## acid            73
## became          73
## focus           73
## experience      72
## face            72
## clear           72
## around          71
## condition       71
## panic           70
## function        70
## several         70
## helps           69
## level           69
## morning         69
## cholesterol     69
## heart           67
## quickly         67
## asleep          66
## significantly   66
## though          65
## make            65
## attacks         64
## know            63
## cold            63
## per             63
## problem         63
## see             63
## although        62
## increase        62
## eliminated      62
## three           61
## works           61
## eyes            61
## bad             61
## immediately     60
## headache        60
## usually         59
## difference      59
## caused          59
## worse           58
## flashes         58
## starting        57
## stomach         57
## ability         57
```

```
## bed             56
## reflux          56
## migraines       56
## drugs           56
## dry             56
## allowed         56
## another         55
## new             55
## gave            55
## body            54
## easy            54
## smoking         53
## last            52
## lowered         52
## enough          51
## etc             51
```

```
# Me quedo solo con los términos
terms_benefits_train_corpus_200 <- rownames(terms_frecuency_benefits_train_corpus_200)
terms_benefits_train_corpus_200
```

```
##   [1] "pain"          "taking"        "drug"          "also"
##   [5] "day"           "skin"          "sleep"         "treatment"
##   [9] "able"          "medication"    "time"          "better"
##  [13] "effects"       "take"          "much"          "feel"
##  [17] "get"           "symptoms"      "depression"    "helped"
##  [21] "less"          "side"          "years"         "reduced"
##  [25] "benefits"      "days"          "acne"          "first"
##  [29] "felt"          "without"       "anxiety"       "life"
##  [33] "within"        "still"         "like"          "now"
##  [37] "effective"     "took"          "started"       "one"
##  [41] "work"          "back"          "months"        "night"
##  [45] "weeks"         "blood"         "well"          "can"
##  [49] "stopped"       "severe"        "just"          "used"
##  [53] "even"          "normal"        "two"           "taken"
##  [57] "feeling"       "help"          "improved"      "mood"
##  [61] "hours"         "increased"     "made"          "pressure"
##  [65] "control"       "good"          "didnt"         "went"
##  [69] "hair"          "cleared"       "energy"        "week"
##  [73] "really"        "use"           "completely"    "daily"
##  [77] "almost"        "weight"        "prescribed"    "reduction"
##  [81] "doctor"        "infection"     "dose"          "away"
##  [85] "decreased"     "times"         "due"           "relief"
##  [89] "none"          "long"          "however"       "year"
##  [93] "will"          "since"         "every"         "effect"
##  [97] "dont"          "little"        "longer"        "patient"
## [101] "loss"          "never"         "worked"        "noticed"
## [105] "seemed"        "lower"         "benefit"       "using"
## [109] "experienced"   "month"         "headaches"     "tried"
## [113] "think"         "gone"          "got"           "high"
## [117] "period"        "medicine"      "overall"       "pill"
## [121] "ive"           "getting"       "great"         "stop"
## [125] "hot"           "began"         "many"          "migraine"
## [129] "mild"          "improvement"   "reduce"        "decrease"
## [133] "things"        "levels"        "dosage"        "problems"
```

```
## [137] "found"         "going"         "results"       "need"
## [141] "acid"          "became"        "focus"         "experience"
## [145] "face"          "clear"         "around"        "condition"
## [149] "panic"         "function"      "several"       "helps"
## [153] "level"         "morning"       "cholesterol"   "heart"
## [157] "quickly"       "asleep"        "significantly" "though"
## [161] "make"          "attacks"       "know"          "cold"
## [165] "per"           "problem"       "see"           "although"
## [169] "increase"      "eliminated"    "three"         "works"
## [173] "eyes"          "bad"           "immediately"   "headache"
## [177] "usually"       "difference"    "caused"        "worse"
## [181] "flashes"       "starting"      "stomach"       "ability"
## [185] "bed"           "reflux"        "migraines"     "drugs"
## [189] "dry"           "allowed"       "another"       "new"
## [193] "gave"          "body"          "easy"          "smoking"
## [197] "last"          "lowered"       "enough"        "etc"
```

Como ya sabemos las palabras a usar, es decir, los 100 términos que más se repite, procedemos a la agrupación por sinónimos. En donde, mediante la función *distance(...)* de la libería *rword2vec*, obtendremos todas palabras más similares de nuestro conjunto, en nuestro caso nos vamos a quedar con las 2 primeras:

```
# http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/
# https://github.com/mukul13/rword2vec
# http://www.rpubs.com/mukul13/rword2vec

library(devtools)                    # hace falta esta librería para que funcione
install_github("mukul13/rword2vec") # nos instalamos la libreria desde Github
```

```
## Skipping install of 'rword2vec' from a github remote, the SHA1 (9942d70f) has not changed since last
##   Use `force = TRUE` to force installation
```

```
library(rword2vec)

# Escribo en un fichero la columna "benefitsReview"
write.table(datos_train$benefitsReview, "benefitsReview.txt", sep = "\t", quote = F, row.names = F)

# Entreno los datos del texto para obtener los vectores de palabras
model_benefits_train = word2vec(train_file = "benefitsReview.txt", output_file = "benefitsReview.bin",
```

```
## Starting training using file benefitsReview.txt
## 100K
Vocab size: 2314
## Words in train file: 100690
```

```
dist_terms_benefits_train_corpus_200 = c()
# Obtengo la distancia de las 100 palabras con mayor frecuencia
for (i in 1:length(terms_benefits_train_corpus_200)){ # calculamos la distancia de la palabra a sus sin
  dist_terms_benefits_train_corpus_200[i] = distance(file_name = "benefitsReview.bin", search_word = te
}
```

```
## Entered word or sentence: pain
##
## Word: pain  Position in vocabulary: 29
```

```
## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: taking
##
## Word: taking  Position in vocabulary: 28

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: drug
##
## Word: drug  Position in vocabulary: 38

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: also
##
## Word: also  Position in vocabulary: 41

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: day
##
## Word: day  Position in vocabulary: 74

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: skin
##
## Word: skin  Position in vocabulary: 50

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: sleep
##
## Word: sleep  Position in vocabulary: 60

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: treatment
##
## Word: treatment  Position in vocabulary: 61

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: able
##
## Word: able  Position in vocabulary: 42

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
```

```
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: medication
##
## Word: medication  Position in vocabulary: 66

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: time
##
## Word: time  Position in vocabulary: 69

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: better
##
## Word: better  Position in vocabulary: 70

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: effects
##
## Word: effects  Position in vocabulary: 73

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: take
##
## Word: take  Position in vocabulary: 48

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: much
##
## Word: much  Position in vocabulary: 59

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: feel
##
## Word: feel  Position in vocabulary: 52

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: get
##
```

```
## Word: get  Position in vocabulary: 51

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: symptoms
##
## Word: symptoms  Position in vocabulary: 72

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: depression
##
## Word: depression  Position in vocabulary: 96

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: helped
##
## Word: helped  Position in vocabulary: 62

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: less
##
## Word: less  Position in vocabulary: 64

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: side
##
## Word: side  Position in vocabulary: 54

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: years
##
## Word: years  Position in vocabulary: 98

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: reduced
##
## Word: reduced  Position in vocabulary: 78

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: benefits
##
## Word: benefits  Position in vocabulary: 85

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: days
##
## Word: days  Position in vocabulary: 99

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: acne
##
## Word: acne  Position in vocabulary: 93

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: first
##
## Word: first  Position in vocabulary: 67

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: felt
##
## Word: felt  Position in vocabulary: 65

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: without
##
## Word: without  Position in vocabulary: 71

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: anxiety
##
## Word: anxiety  Position in vocabulary: 110

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: life
##
## Word: life  Position in vocabulary: 123

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
```

```
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: within
##
## Word: within  Position in vocabulary: 83

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: still
##
## Word: still  Position in vocabulary: 75

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: like
##
## Word: like  Position in vocabulary: 76

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: now
##
## Word: now  Position in vocabulary: 112

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: effective
##
## Word: effective  Position in vocabulary: 97

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: took
##
## Word: took  Position in vocabulary: 80

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: started
##
## Word: started  Position in vocabulary: 84

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: one
##
```

```
## Word: one  Position in vocabulary: 94

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: work
##
## Word: work  Position in vocabulary: 111

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: back
##
## Word: back  Position in vocabulary: 102

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: months
##
## Word: months  Position in vocabulary: 114

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: night
##
## Word: night  Position in vocabulary: 135

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: weeks
##
## Word: weeks  Position in vocabulary: 130

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: blood
##
## Word: blood  Position in vocabulary: 91

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: well
##
## Word: well  Position in vocabulary: 139

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: can
##
## Word: can  Position in vocabulary: 92

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: stopped
##
## Word: stopped  Position in vocabulary: 119

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: severe
##
## Word: severe  Position in vocabulary: 107

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: just
##
## Word: just  Position in vocabulary: 103

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: used
##
## Word: used  Position in vocabulary: 106

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: even
##
## Word: even  Position in vocabulary: 115

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: normal
##
## Word: normal  Position in vocabulary: 141

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: two
##
## Word: two  Position in vocabulary: 109

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
```

```
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: taken
##
## Word: taken  Position in vocabulary: 117

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: feeling
##
## Word: feeling  Position in vocabulary: 113

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: help
##
## Word: help  Position in vocabulary: 120

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: improved
##
## Word: improved  Position in vocabulary: 149

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: mood
##
## Word: mood  Position in vocabulary: 161

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: hours
##
## Word: hours  Position in vocabulary: 166

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: increased
##
## Word: increased  Position in vocabulary: 138

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: made
##
```

```
## Word: made  Position in vocabulary: 121

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: pressure
##
## Word: pressure  Position in vocabulary: 131

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: control
##
## Word: control  Position in vocabulary: 154

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: good
##
## Word: good  Position in vocabulary: 140

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: didnt
##
## Word: didnt  Position in vocabulary: 1007

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: went
##
## Word: went  Position in vocabulary: 122

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: hair
##
## Word: hair  Position in vocabulary: 137

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: cleared
##
## Word: cleared  Position in vocabulary: 136

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: energy
##
## Word: energy  Position in vocabulary: 185

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: week
##
## Word: week  Position in vocabulary: 187

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: really
##
## Word: really  Position in vocabulary: 127

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: use
##
## Word: use  Position in vocabulary: 150

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: completely
##
## Word: completely  Position in vocabulary: 168

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: daily
##
## Word: daily  Position in vocabulary: 156

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: almost
##
## Word: almost  Position in vocabulary: 134

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: weight
##
## Word: weight  Position in vocabulary: 153

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
```

```
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: prescribed
##
## Word: prescribed  Position in vocabulary: 148

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: reduction
##
## Word: reduction  Position in vocabulary: 155

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: doctor
##
## Word: doctor  Position in vocabulary: 152

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: infection
##
## Word: infection  Position in vocabulary: 180

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: dose
##
## Word: dose  Position in vocabulary: 167

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: away
##
## Word: away  Position in vocabulary: 208

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: decreased
##
## Word: decreased  Position in vocabulary: 188

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: times
##
```

```
## Word: times  Position in vocabulary: 163

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: due
##
## Word: due  Position in vocabulary: 145

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: relief
##
## Word: relief  Position in vocabulary: 178

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: none
##
## Word: none  Position in vocabulary: 345

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: long
##
## Word: long  Position in vocabulary: 162

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: however
##
## Word: however  Position in vocabulary: 451

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: year
##
## Word: year  Position in vocabulary: 233

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: will
##
## Word: will  Position in vocabulary: 147

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: since
##
## Word: since  Position in vocabulary: 199

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: every
##
## Word: every  Position in vocabulary: 157

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: effect
##
## Word: effect  Position in vocabulary: 182

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: dont
##
## Word: dont  Position in vocabulary: 1019

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: little
##
## Word: little  Position in vocabulary: 165

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: longer
##
## Word: longer  Position in vocabulary: 160

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: patient
##
## Word: patient  Position in vocabulary: 193

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: loss
##
## Word: loss  Position in vocabulary: 226

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
```

```
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: never
##
## Word: never  Position in vocabulary: 158

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: worked
##
## Word: worked  Position in vocabulary: 181

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: noticed
##
## Word: noticed  Position in vocabulary: 159

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: seemed
##
## Word: seemed  Position in vocabulary: 169

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: lower
##
## Word: lower  Position in vocabulary: 175

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: benefit
##
## Word: benefit  Position in vocabulary: 190

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: using
##
## Word: using  Position in vocabulary: 171

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: experienced
##
```

```
## Word: experienced  Position in vocabulary: 191

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: month
##
## Word: month  Position in vocabulary: 244

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: headaches
##
## Word: headaches  Position in vocabulary: 249

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: tried
##
## Word: tried  Position in vocabulary: 186

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: think
##
## Word: think  Position in vocabulary: 170

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: gone
##
## Word: gone  Position in vocabulary: 332

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: got
##
## Word: got  Position in vocabulary: 179

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: high
##
## Word: high  Position in vocabulary: 222

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

```
## Entered word or sentence: period
##
## Word: period  Position in vocabulary: 241

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: medicine
##
## Word: medicine  Position in vocabulary: 254

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: overall
##
## Word: overall  Position in vocabulary: 234

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: pill
##
## Word: pill  Position in vocabulary: 252

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: ive
##
## Word: ive  Position in vocabulary: -1
## Out of dictionary word!

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: getting
##
## Word: getting  Position in vocabulary: 197

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: great
##
## Word: great  Position in vocabulary: 255

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: stop
##
## Word: stop  Position in vocabulary: 202
```

```
## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: hot
##
## Word: hot  Position in vocabulary: 210

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: began
##
## Word: began  Position in vocabulary: 195

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: many
##
## Word: many  Position in vocabulary: 204

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: migraine
##
## Word: migraine  Position in vocabulary: 236

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: mild
##
## Word: mild  Position in vocabulary: 219

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: improvement
##
## Word: improvement  Position in vocabulary: 212

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: reduce
##
## Word: reduce  Position in vocabulary: 205

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: decrease
```

```
##
## Word: decrease  Position in vocabulary: 221

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: things
##
## Word: things  Position in vocabulary: 224

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: levels
##
## Word: levels  Position in vocabulary: 275

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: dosage
##
## Word: dosage  Position in vocabulary: 257

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: problems
##
## Word: problems  Position in vocabulary: 285

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: found
##
## Word: found  Position in vocabulary: 201

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: going
##
## Word: going  Position in vocabulary: 214

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: results
##
## Word: results  Position in vocabulary: 286

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
```

```
## múltiplo de la longitud del reemplazo

## Entered word or sentence: need
##
## Word: need  Position in vocabulary: 200

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: acid
##
## Word: acid  Position in vocabulary: 223

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: became
##
## Word: became  Position in vocabulary: 206

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: focus
##
## Word: focus  Position in vocabulary: 277

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: experience
##
## Word: experience  Position in vocabulary: 227

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: face
##
## Word: face  Position in vocabulary: 279

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: clear
##
## Word: clear  Position in vocabulary: 247

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: around
##
## Word: around  Position in vocabulary: 216
```

```
## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: condition
##
## Word: condition  Position in vocabulary: 370

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: panic
##
## Word: panic  Position in vocabulary: 228

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: function
##
## Word: function  Position in vocabulary: 270

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: several
##
## Word: several  Position in vocabulary: 218

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: helps
##
## Word: helps  Position in vocabulary: 263

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: level
##
## Word: level  Position in vocabulary: 292

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: morning
##
## Word: morning  Position in vocabulary: 341

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: cholesterol
```

```
##
## Word: cholesterol  Position in vocabulary: 264

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: heart
##
## Word: heart  Position in vocabulary: 235

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: quickly
##
## Word: quickly  Position in vocabulary: 290

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: asleep
##
## Word: asleep  Position in vocabulary: 320

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: significantly
##
## Word: significantly  Position in vocabulary: 356

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: though
##
## Word: though  Position in vocabulary: 300

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: make
##
## Word: make  Position in vocabulary: 231

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: attacks
##
## Word: attacks  Position in vocabulary: 350

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
```

```
## múltiplo de la longitud del reemplazo

## Entered word or sentence: know
##
## Word: know  Position in vocabulary: 239

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: cold
##
## Word: cold  Position in vocabulary: 240

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: per
##
## Word: per  Position in vocabulary: 229

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: problem
##
## Word: problem  Position in vocabulary: 405

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: see
##
## Word: see  Position in vocabulary: 258

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: although
##
## Word: although  Position in vocabulary: 382

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: increase
##
## Word: increase  Position in vocabulary: 245

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: eliminated
##
## Word: eliminated  Position in vocabulary: 308
```

```
## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: three
##
## Word: three  Position in vocabulary: 246

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: works
##
## Word: works  Position in vocabulary: 271

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: eyes
##
## Word: eyes  Position in vocabulary: 504

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: bad
##
## Word: bad  Position in vocabulary: 281

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: immediately
##
## Word: immediately  Position in vocabulary: 364

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: headache
##
## Word: headache  Position in vocabulary: 288

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: usually
##
## Word: usually  Position in vocabulary: 267

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: difference
```

```
##
## Word: difference  Position in vocabulary: 325

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: caused
##
## Word: caused  Position in vocabulary: 250

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: worse
##
## Word: worse  Position in vocabulary: 447

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: flashes
##
## Word: flashes  Position in vocabulary: 371

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: starting
##
## Word: starting  Position in vocabulary: 272

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: stomach
##
## Word: stomach  Position in vocabulary: 348

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: ability
##
## Word: ability  Position in vocabulary: 280

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: bed
##
## Word: bed  Position in vocabulary: 354

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
```

```
## múltiplo de la longitud del reemplazo

## Entered word or sentence: reflux
##
## Word: reflux  Position in vocabulary: 390

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: migraines
##
## Word: migraines  Position in vocabulary: 399

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: drugs
##
## Word: drugs  Position in vocabulary: 378

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: dry
##
## Word: dry  Position in vocabulary: 314

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: allowed
##
## Word: allowed  Position in vocabulary: 297

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: another
##
## Word: another  Position in vocabulary: 302

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: new
##
## Word: new  Position in vocabulary: 289

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: gave
##
## Word: gave  Position in vocabulary: 266
```

```
## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: body
##
## Word: body  Position in vocabulary: 359

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: easy
##
## Word: easy  Position in vocabulary: 358

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: smoking
##
## Word: smoking  Position in vocabulary: 389

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: last
##
## Word: last  Position in vocabulary: 278

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: lowered
##
## Word: lowered  Position in vocabulary: 310

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: enough
##
## Word: enough  Position in vocabulary: 322

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo

## Entered word or sentence: etc
##
## Word: etc  Position in vocabulary: 1493

## Warning in dist_terms_benefits_train_corpus_200[i] <- distance(file_name
## = "benefitsReview.bin", : número de items para para sustituir no es un
## múltiplo de la longitud del reemplazo
```

Una vez, que tenemos todas las palabras con los 3 términos más similares, procedemos a sustituir todos esos términos por el término general, es decir:

```
# Obtenemos el primer término -> "pain"
terms_benefits_train_corpus_200[1]
```

```
## [1] "pain"
```

```
# Vamos a sustituir "pain" por sus dos palabras más similares
dist_terms_benefits_train_corpus_200[[1]]
```

```
## [1] nausea relief
## Levels: nausea relief
```

Por último, ya solo nos queda hacer el reemplazamiento, para ello se usará la función *gsub(...)* sobre el corpus (benefits_corpus). Para sustituir las palabras en el texto, se ha uso de la función `gsub(pattern, replacement, x, ignore.case = FALSE, perl = FALSE, fixed = FALSE, useBytes = FALSE)`.

```
# Para la columna benefitsReview del conjunto train

for (i in 1:10) # iteramos sobre los terminos # SE DEBE CAMBIAR PARA VER CUANTOS TÉRMINOS COGEMO
  for (j in 1:2) # iteramos sobre los sinónimos, en este caso solo tenemos 2
    benefits_train_corpus_new <- tm_map(benefits_train_corpus, content_transformer(gsub),
                                        pattern = as.character(dist_terms_benefits_train_corpus_200[[i]][j]),
                                        replacement = as.character(terms_benefits_train_corpus_200[i]))

# Comprobamos que efectivamente se han producido cambios, por ejemplo al revisar el término "medication
write.table(benefits_train_corpus$content, "sinSinonimos.txt")
write.table(benefits_train_corpus_new$content, "conSinonimos.txt")
```

2.2.8. Stemming

El siguiente paso consiste en reducir el número de palabras totales con las que estamos trabajando. En este caso, se trata de reducir aquellas que no nos aportan nada relevante a lo que ya tenemos. En la columna con la que estamos trabajando en este dataframe, se repite una gran cantidad de veces la palabra "benefit", al igual que "benefits".

Sin embargo, realizar el análisis de nuestros datos con ambas palabras no tiene gran relevancia, ya que una no aporta nada respecto a la otra. Este es un ejemplo del tipo de casos que se nos dan en nuestro dataset. Igual ocurre con "reduce" y "reduced", por ejemplo. Este tipo de situaciones son las que intentamos corregir con este paso. Vamos a ver un ejemplo de este suceso, que se da por ejemplo en los siguientes valores del corpus (y en muchos más).

```
inspect(benefits_train_corpus_new[183])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
## [1]  treatment benefits   temporary    made sneezing  watery eyes diminish    address  issue  respira
```

```
inspect(benefits_train_corpus_new[213])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
## [1] overall   ease mantally    true benefit  felt
```

A continuación, aplicamos el proceso de stemming mediante la siguiente orden:

```
benefits_train_corpus_new <- tm_map(benefits_train_corpus_new, stemDocument)
#effects_train_corpus_new <- tm_map(effects_train_corpus_new, stemDocument)

#benefits_test_corpus_new <- tm_map(benefits_test_corpus_new, stemDocument)
#effects_test_corpus_new <- tm_map(effects_test_corpus_new, stemDocument)
```

Si ahora volvemos a mostrar el contenido de dichas opiniones, podemos ver que el stemming se ha hecho efectivo: donde ponía *benefits*, ahora pone *benefit*, como se puede comprobar si volvemos a mostrar dichos elementos del corpus. De hecho, si nos fijamos, no solo esta palabra ha resultado modificada, sino que se han resumido muchas más palabras en comparación a como teníamos los documentos en el momento previo a la aplicación del método *Stem*. Desde este momento, ya tenemos nuestro conjunto reducido a nivel de concepto.

```
inspect(benefits_train_corpus_new[183])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
## [1] treatment benefit temporari made sneez wateri eye diminish address issu respiratori difficulti f
```

```
inspect(benefits_train_corpus_new[213])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
## [1] overal eas mantal true benefit felt
```

2.2.9. Borrar espacios en blanco innecesarios

Hasta el momento hemos hecho distintos cambios en el texto de nuestro dataset. No solo hemos modificado algunas palabras, sino que también hemos borrado otras muchas. Por ello, es adecuado asegurarnos que no hay más espacios en blanco que los que separan las palabras del texto. Para asegurarnos de ello, podemos ejecutar la siguiente orden, que se encarga de suprimir los espacios en blanco sobrantes.

```
benefits_train_corpus_new <- tm_map(benefits_train_corpus_new, stripWhitespace)
#effects_train_corpus_new <- tm_map(effects_train_corpus_new, stripWhitespace)

#benefits_test_corpus_new <- tm_map(benefits_test_corpus_new, stripWhitespace)
#effects_test_corpus_new <- tm_map(effects_test_corpus_new, stripWhitespace)
```

2.2.10. Term Document Matrix

Ahora vamos a mapear nuestro corpus creando una matriz de términos, donde las filas corresponden a los documentos y las columnas a los términos. Para ello usaremos la función TermDocumentMatrix:

```
matrix_corpus <- TermDocumentMatrix(benefits_train_corpus_new)
```

Podemos observar que tenemos 5838 términos, esto quiere decir que tenemos 5838 palabras diferentes en nuestro Corpus. Obtengamos la *frecuencia de las palabras*:

```
class(matrix_corpus)
```

```
## [1] "TermDocumentMatrix"    "simple_triplet_matrix"
```

Como podemos ver, actualmente aún no tenemos nuestros datos en la matriz que buscamos, sino en un vector, por tanto:

```
matrix_corpus <- as.matrix(matrix_corpus)
class(matrix_corpus)
```

```
## [1] "matrix"
```

```
dim(matrix_corpus)
```

```
## [1] 5837 3107
```

Con este método, hemos obtenido la ocurrencia de las palabras que tenemos en nuestro dataset para cada uno de los documentos/comentarios. Esta matriz tiene 5838 columnas, que representa la totalidad de palabras diferentes que hay en los comentarios de la columna *benefitsReview*, y 3107 filas, donde cada una representa un comentario. Por tanto, en la fila iésima la matriz, tendremos la ocurrencia de las palabras en *benefitsReview* que existen en el comentario *i*.

```
# Sumamos las filas
suma_matrix_corpus <- rowSums(matrix_corpus)
head(suma_matrix_corpus,5)
```

```
##    agent     alon  congest dysfunct   failur
##        2       19       19        2        7
```

```
# Ordenanamos de mayor a menor y muestra los 10 primeros
ordena_mayor_matrix_corpus <- sort(suma_matrix_corpus, decreasing = TRUE)
head(ordena_mayor_matrix_corpus,10)
```

```
##   take effect   pain    day   help   drug  medic   feel   time   work
##    789    682    643    626    524    498    488    479    450    404
```

```
copia_ordena_mayor = ordena_mayor_matrix_corpus # Para graficos (evitando data.frame)

# Ordenanamos de menor a mayor y muestra los 10 primeros
ordena_menor_matrix_corpus <- sort(suma_matrix_corpus, decreasing = FALSE)
head(ordena_menor_matrix_corpus,10)
```

```
##         mangag          overt    ventricular            con           pros
##              1              1              1              1              1
##        ponstel          frank       valerian allergiesirrit          dryer
##              1              1              1              1              1
```

```
# Transformamos a objeto data.frame, con dos columnas (palabra, frec), para posteriormente graficarlo.
ordena_mayor_matrix_corpus <- data.frame(palabra = names(ordena_mayor_matrix_corpus), frec = ordena_may
```

Creamos la nube de palabras:

```
# instalar paquete worldcloud

#wordcloud(
#  words = ordena_mayor_matrix_corpus$palabra,
#  freq = ordena_mayor_matrix_corpus$frec,
#  max.words = 80,
#  random.order = F,
#  colors=brewer.pal(name = "Dark2", n = 8)
#  )
```

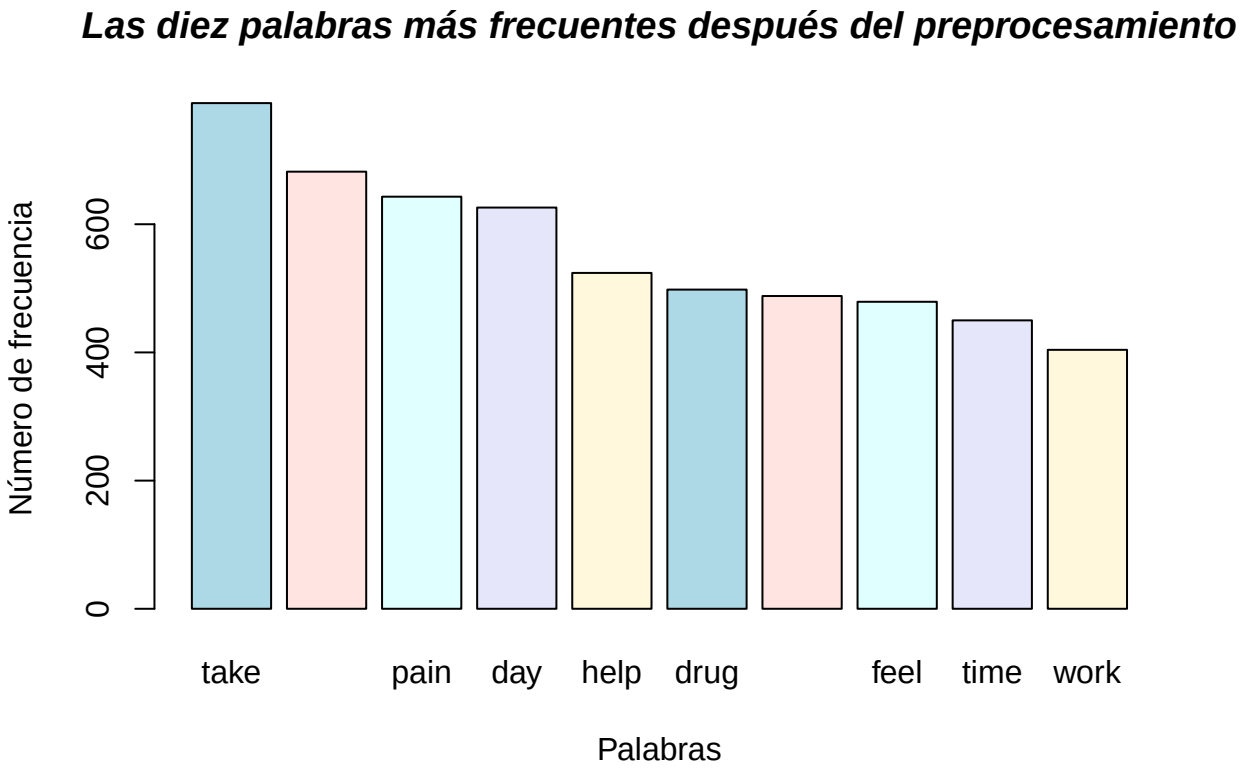
Mostramos las más frecuentes:

```
ordena_mayor_matrix_corpus[1:20,]
```

```
##             palabra frec
## take           take  789
## effect       effect  682
## pain           pain  643
## day             day  626
## help           help  524
## drug           drug  498
## medic         medic  488
## feel           feel  479
## time           time  450
## work           work  404
## also           also  394
## use             use  381
## sleep         sleep  381
## reduc         reduc  381
## year           year  369
## get             get  368
## skin           skin  358
## treatment treatment  357
## benefit     benefit  353
## depress     depress  349
```

Y obtenemos la *gráfica*:

```
copia_ordena_mayor <- as.matrix(copia_ordena_mayor)
barplot(copia_ordena_mayor[1:10,],  xlab="Palabras", ylab="Número de frecuencia",
        col = c("lightblue", "mistyrose", "lightcyan",
                "lavender", "cornsilk"))
title(main = list("Las diez palabras más frecuentes después del preprocesamiento", font = 4))
```


Las diez palabras más frecuentes después del preprocesamiento
Número de frecuencia
0
200
400
600
take
pain
day
help
drug
feel
time
work
Palabras