Lo primero que se debe hacer es descargar cada uno de los planes de desarrollo municipal para los municipios de interés. Dado que los planes de desarrollo municipal vienen en formato de documento portátil (PDF) es necesario copiarlo a un bloc de notas y posteriormente guardarlo (es importante mencionar que en ningún caso se tuvo en cuenta los índices del documento).

**Comenzando**

Lo primero que se necesita es contar con el software R el cual puede ser descargado de manera gratuita de la página: <https://cran.r-project.org/>

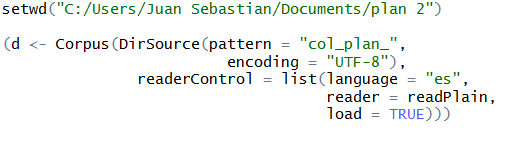
Una vez se descarga el software se deben instalar dos librerías: Wordcloud y TM. La descripción de cada uno de los paquetes puede ser vista en

<https://cran.r-project.org/web/packages/tm/tm.pdf>

<https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>



Una vez se cargan las librerías se procede a fijar el directorio y cargar los documentos.

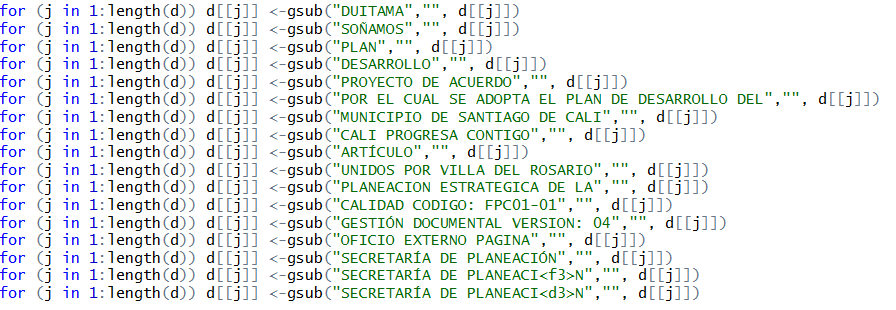


Con el comando setwd se fija el directorio y con el comando siguiente lo que se hace es cargar los documentos que están en formato de texto y que además empiezan con el patrón “col\_plan\_” por ejemplo en nuestro caso cada plan de desarrollo tenía un nombre asi col\_plan\_bogota. De esta forma el programa cargará todos los planes de desarrollo de una sola vez.

Lo que se hace a continuación es quitar los espacios en blanco de los planes de desarrollo.



Luego es necesario limpiar cada uno de los archivos al máximo eliminando expresiones que quedaran pegadas del PDF.

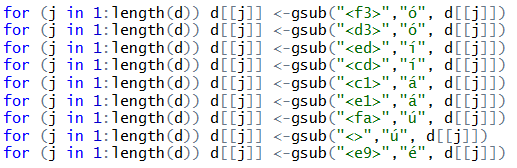


Para lograr este trabajo de limpieza utilizamos el comando gsub que se utiliza para buscar y remplazar. De esta forma lo que hace la macro es buscar las expresiones que se tienen allí y eliminarlas en cada uno de los planes de desarrollo.

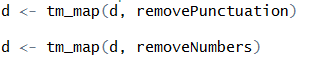
El siguiente paso es llevar el texto de cada uno de los planes de desarrollo a minúsculas.



El paquete “tm” no reconoce las tildes que tienen las palabras así que se debe solucionar este error de la siguiente manera



Posteriormente se elimina la puntuación y los números



Se realiza una agrupación de términos para poder tener una medida más exacta del uso de TIC en los planes de desarrollo. La agrupación se hizo de la siguiente manera:

**Interconectado:** interconectada, interconectados, interconectividad, interconexión, interconexiones.

**Productividad:** productiva, productivos, productivo, productivas.

**Sostenible:** Sostenibles, sostenibilidad.

**Optimizar:** Optimización, optimizado, optimizados, optimizando,

**Innovador:** innovadora, innovación, innovaciones, innovadores, innovar, innovando.

**Inteligente:** inteligentes, inteligencia.

**Digital:** digitales, digitalización

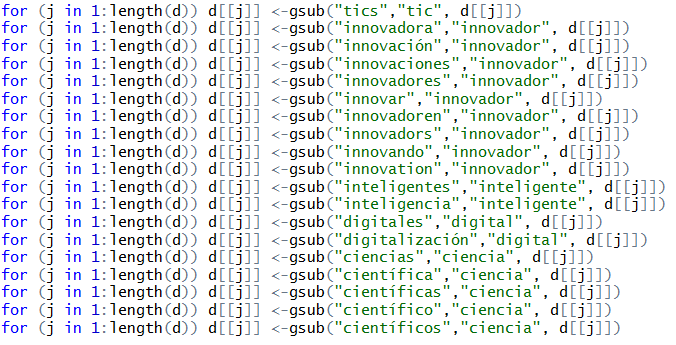
**Ciencia:** ciencias, científica, científicas, científico, científicos

**Tecnología:** tecnológica, tecnologías, tecnología, tecnológicas, tecnológico, tecnológicos, tecnificación, tecnologías.

**Tic:** tics

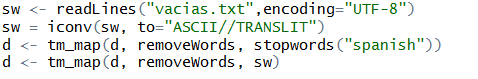
**Conectar:** conectarse, conectividad.

**Competitividad:** competitiva, competitivas, competitivo, competitivos.



Un pequeño ejemplo de cómo se agrupó es el anterior.

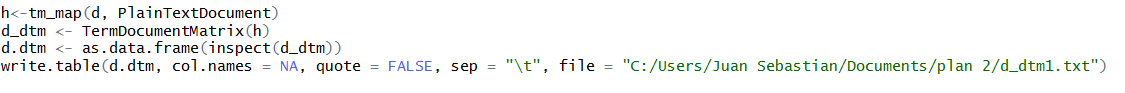
Por último se deben remover palaras que no tienen sentido ser analizadas como por ejemplo las preposiciones. El listado completo de palabras que fueron eliminados de los textos se encuentra adjunto.



El comando readLines sirve para cargar un archivo, en este caso se carga el diccionario con parte de las palabras que van a ser eliminadas. La otra parte que son removidas viene de usar el comando stopwords que contiene otras palabras que no tienen sentido.

**Construcción de tabla de frecuencias**

Para construir la tabla de frecuencias basta con hacerle una pequeña transformación a lo que anterior. Esto es:



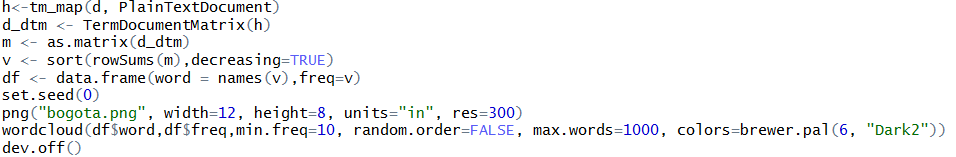
Se toman los documentos que han quedado luego de hacer toda la limpieza, se convierten en formato “PlaintTextDocument” el cual permite ya hacer el conteo de palabras pues sobre este es que se usa el comando TermDocumentMatrix que hace una matriz contando el número de veces que se repite cada una de las palabras en el texto. Por último se cambia a formato data.frame que es el formato de base de datos y se exporta en formato de texto para abrirlo en Excel y manipularlo con el comando write.table.

**Construcción de nubes de palabras**

Para la construcción de nubes de palabras es necesario usar el anterior código pero cargando documento por documento, es decir no se cargan todos los planes de desarrollo de una sala vez sino uno por uno. Para esto lo único que debe cambiarse de lo explicado anteriormente es que en la parte donde se escribe el patrón de carga de documentos, cambiándolo por el nombre exacto de esta manera por ejemplo para el caso para Bogotá se reemplaza por col\_plan\_bogota. También se agregan los términos para poder ver que tanto se nombra las palabras tic y relacionadas con tic en el documento. Las agrupaciones son:

Tic: innovador, productividad.

Luego se usa el mismo código pero para generar la nube de datos:



Se convierte nuevamente a formato “PlainTextDocument” y las siguientes líneas generan la matriz de frecuencias pero ordenada de manera decreciente en cuanto al número de repeticiones. El comando set.seed funciona para generar la misma nube siempre (puesto que el giro de las palabras puede ser diferente por ejemplo) y las tres últimas líneas son las que generan la nube de palabras en este caso se tienen en cuenta palabras con mínimo 10 repeticiones y con un numero de máximo 1000 palabras en la nube. Un ejemplo es el siguiente:

