Word Cloud using R

Kanksha Masrani

14 July 2018

## Word Cloud

The simplest data visualization is a word cloud. It is one of the Text mining techniques used to visualize the most frequent occurring words in a given file (.csv/.text). The more frequent the word is used, the larger and bolder it is displayed.Text mining refers to the process of deriving high quality information from text.

The aim of this article is to explain the concept of Word Cloud and understand how to actually create a Word Cloud using R.

## 1. Choosing the Text File :

Choose the text you wish to create a word cloud out of. For example here I am going to create a word cloud from the transcript of a House of Lords debate. Copy and paste the text into a plain text file (word.txt). You can also import a .csv file.

## 2. Installing Packages:

You will need to install and load the following packages. We will require four packages.

1. Tm- Text mining package
2. SnowballC- Text stemming
3. wordCloud- Wordcloud package
4. RcolorBrewer- Color palettes for graphics

library("tm")  
library("SnowballC")  
library("wordcloud")  
library("RColorBrewer")

## 3. Reading the file:

Following are the command to read a text file or a csv file in R

speech ="C:\\Users\\Admin\\Desktop\\Word.txt"  
text = readLines(speech)  
  
##speech ="C:\\Users\\Admin\\Desktop\\Word.csv"  
##text = readLines(speech)

## 4. Converting the text/.csv file into a corpus:

Now in order to process or clean the text using tm package, you need to first convert this plain text data into a format called corpus which can then be processed by the tm package. A corpus is a collection of documents (although in our case we only have one).

Following is the command to convert .txt file into a corpus.

docs <- Corpus(VectorSource(text))

## 5. Data Cleaning:

Execute the following commands in RStudio:

inspect(docs)

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 11  
##   
## [1] There is some anxiety among those involved with sickle cell services that the complexity of the services needed effectively places them largely outside the scope of the clinical commissioning groups. Many are concerned about the type of policies that will be in place to ensure that a patient-centred, integrated approach to care engages primary care and community interests across health, social and community care. This is to help to reduce morbidity, needless hospital care and the health inequalities experienced by this seriously marginalised sector.   
## [2]   
## [3] There are expectations that not only CCGs but local health and well-being boards should aim to reflect the make-up of their respective client communities. So, given that the steady establishment of CCGs and the view that community provision of sickle cell disorder management have a major role to play across the country, especially in high-risk areas within CCGs, can the Minister tell the House what priority is being given by CCGs to people in the sickle cell and thalassaemia community, who are feeling concerned, vulnerable and anxious about the situation and their future?   
## [4]   
## [5] As yet, there is no cure for sickle cell and more research is needed both for a cure and for the treatment of current sufferers. The existing treatment involves a form of chemotherapy, which can have harmful side effects, such as damage to the immune system. Fortunately, Sparks, a charity which provides funding into research for childhood diseases<U+0097>I declare an interest as a trustee<U+0097>is funding a research project that aims to investigate the possibility of a safer, less toxic and more targeted therapy. However, in the mean time, there needs to be widespread education and awareness among those who assess the level of disability of sickle cell sufferers. They need to be made more aware and educated about the situation faced by people living with sickle cell and its associated conditions.  
## [6]   
## [7] The Government also need to seriously improve the awareness of the wider population about the plight of people living with this inherited blood disorder and the disabilities that they may be facing, quite often invisibly so.   
## [8]   
## [9] I know that the Sickle Cell Society, the UK Thalassaemia Society and the UK Forum on Haemoglobin Disorders would be more than willing to meet the appropriate government departments and agencies to discuss how they can work together to address the serious concerns that I have highlighted. I hope that this offer will be acted upon.   
## [10]   
## [11] As the last US election showed, BME communities vote for people who they consider address their needs and concerns. This should be food for thought for us on this side of the Atlantic. I look forward to hearing my noble friend<U+0092>s response, as I know that she is always sympathetic to inequality issues and, like me, strives towards a just and fair society.

toSpace <- content\_transformer(function (x , pattern ) gsub(pattern, " ", x))  
docs <- tm\_map(docs, toSpace, "/")  
docs <- tm\_map(docs, toSpace, "@")  
docs <- tm\_map(docs, toSpace, "\\|")  
# Convert the text to lower case  
docs <- tm\_map(docs, content\_transformer(tolower))  
# Remove numbers  
docs <- tm\_map(docs, removeNumbers)  
# Remove english common stopwords  
docs <- tm\_map(docs, removeWords, stopwords("english"))  
# Remove your own stop word  
# specify your stopwords as a character vector  
docs <- tm\_map(docs, removeWords, c("blabla1", "blabla2"))   
# Remove punctuations  
docs <- tm\_map(docs, removePunctuation)  
# Eliminate extra white spaces  
docs <- tm\_map(docs, stripWhitespace)  
# Text stemming  
# docs <- tm\_map(docs, stemDocument)  
  
#docs <- tm\_mp(docs, removeWords,c("noble","lords"))

As you can see the commands above, use tm\_map() from the tm package for processing your text. As the commands are quite obvious, they do the following:

1. Removes unnecessary white space
2. Converts everything to lower case (since tm package is case sensitive)
3. Removes English common words like ‘the’, ‘I’, ‘Me’ (so-called ‘stopwords’)
4. You can also explicitly remove numbers and punctuation with the removeNumbers and removePunctuation arguments.
5. You can also make a list of words. C(“noble”, “lord”, etc..), to remove them altogether.
6. Performs text stemming which means that all the words are converted to their stem. (Ex: walked -> walk; talking -> talk). This helps us to ensure that all the words are converted to the same form.

## 6. Creating a Document-term Matrix:

It is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to words in the collection and columns correspond to documents.

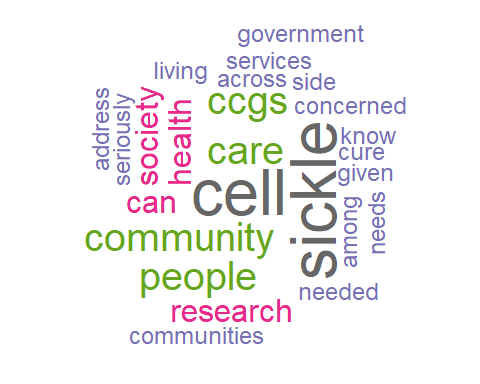
Now we can create a word cloud even without a DTM. But the advantage of using this here is to take a look at the frequency of words.

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

## word freq  
## cell cell 7  
## sickle sickle 7  
## care care 4  
## community community 4  
## ccgs ccgs 4  
## people people 4  
## health health 3  
## can can 3  
## research research 3  
## society society 3

## 7. Your First Word Cloud

wordcloud(words = d$word, freq = d$freq, min.freq = 1,  
 max.words=25, random.order=FALSE, rot.per=0.35,   
 colors=brewer.pal(8, "Dark2"))



Arguments of the word cloud generator function:

1. words : the words to be plotted
2. freq : their frequencies
3. min.freq : words with frequency below min.freq will not be plotted
4. max.words : maximum number of words to be plotted
5. random.order : plot words in random order. If false, they will be plotted in decreasing frequency
6. rot.per : proportion words with 90 degree rotation (vertical text)
7. colors : color words from least to most frequent. Use, for example, colors =“black” for single color.

# Where are they most often used?

1. Marketing: Word Cloud from all the customer reviews of a product. Helps you target profitable customers.
2. Sentiment Analysis: Collect tweets or any data as a matter of fact and you can easily perform sentimental analysis.
3. Distil big topics into bite-sized idea.
4. Summarizing any Poll taken: If you’re asking the audience to vote on, say, 50 items or more, showing 50 bars stacked on one graph isn’t the most. elegant solution. Word Cloud to the rescue!
5. Share latest political Trends.
6. Evaluating you brand Identity

# List of free word cloud generators

1. Tag crowd
2. Make word mosaic
3. Word sift
4. Tagxedo
5. Word clouds
6. Word it out
7. Tableau