**Задание 7 (стр. 46):**

Скачайте тексты трех русских народных сказок. Сформируйте облако слов, используя stopwords("russian"). Какие слова следует исключить? Добавьте эти слова в вектор исключаемых слов и сформируйте новое облако. Продемонстрируйте приемы формирования различной цветовой гаммы.

**Ход работы:**

Для начала необходимо создать R script lab2r.R, в котором будут использованы библиотеки rvest, tm, RColorBrewer, wordcloud которые необходимо предварительно установить командами «install.packages('rvest')» , «install.packages('tm')»,«install.packages('RColorBrewer')», «install.packages('wordcloud')». Эти пакеты нужны будут для работы с «корпусом» документа, составления частотной матрицы слов в документе и облака слов. В качестве исходных текстов были выбраны сказки с сайта nukadeti.ru «Каша из топора», «Маша и Медведь» и «Колобок». Тэг в html документе на котором расположен текст сказок имеет идентификатор "div.tale-text.si-text". После извлечения текстов, объединения их в один вектор res и удаления лишних символов, преобразуем полученный вектор в датафрейм, разбивая текст на отдельные слова. Когда был получен датафрейм со словами в каждой ячейке, можно приступить к удалению стопслов, составлению частотной матрицы и облака слов. Результат работы скрипта приведен на рисунке 1.

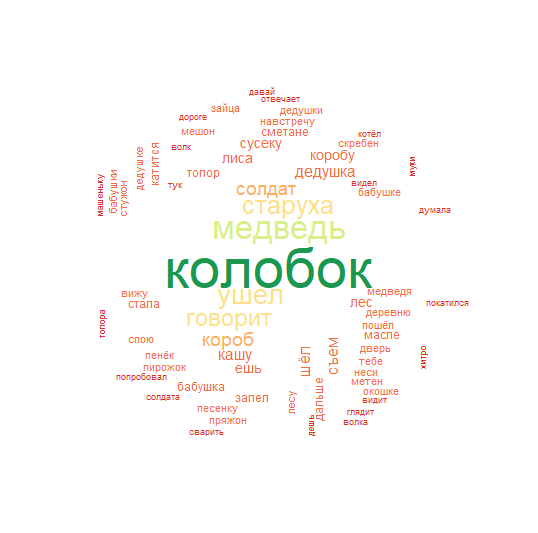


Рисунок 1 – Полученное облако слов

Код файла lab2r.R представлен ниже.

#загрузка требуемых пакетов

require(rvest)

require(tm)

#извлечение из известного узла контента для каждой из 3х сказок

a = "https://nukadeti.ru/skazki/kasha\_iz\_topora"

ar = read\_html(a)

res1 = as.character(html\_text(html\_node(ar, "div.tale-text.si-text")))

b = "https://nukadeti.ru/skazki/masha\_i\_medved"

br = read\_html(b)

res2 = as.character(html\_text(html\_node(br, "div.tale-text.si-text")))

c = "https://nukadeti.ru/skazki/kolobok"

cr = read\_html(c)

res3= as.character(html\_text(html\_node(cr, "div.tale-text.si-text")))

# удаление лишних символов в полученных текстах

res = paste(res1,res2,res3)

res = gsub("\r\n","",res)

res = gsub("[^а-яА-Яё]"," ",res)

res = gsub(" "," ",res)

#преобразование вектора с текстом в фрейм с разделителем пробелом

res = data.frame(strsplit(as.character(res)," "))

names(res)[1] <- "text"

#DataframeSource требует 2х колонок в структуре фрейма а именно id и value

res\_title <- data.frame(doc\_id=row.names(res),

text=res$text)

corpus <- Corpus(DataframeSource(res\_title))

#приведение к нижнему регистру, удаление стопслов, а также удаление “ка”,

#оставшееся от слова дай-ка

corpus <- tm\_map(corpus, tolower)

corpus <- tm\_map(corpus, function(x) removeWords(x, stopwords("russian")))

corpus <- tm\_map(corpus, function(x) removeWords(x, "ка"))

tdm <- TermDocumentMatrix(corpus,control = list(minWordLength = 1))

m <- as.matrix(tdm)

v <- sort(rowSums(m),decreasing=TRUE)

df <- data.frame(word = names(v),freq=v)

require(RColorBrewer)

require(wordcloud)

#установка палитры цветов

pal2 <- brewer.pal(8,"RdYlGn")

#создание облака

wordcloud(df$word,df$freq, scale=c(3.5,0.25),min.freq=3,

max.words=100, random.order=FALSE, rot.per=.15, colors=pal2)