УДК 517.51

АНАЛИЗ НАСТРОЕНИЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ ПРИ ПОМОЩИ СРЕДЫ R

Моложенко Е. С.
Северо-Казахстанский Государственный Университет имени М. Козыбаева, г. Петропавловск

Научный руководитель - к.ф.-м.н., профессор Куликов В.П.

Система статистической обработки данных и программирования R ориентирована на использование интерфейса командной строки. Обработка данных в системе R представляет собой последовательность команд для загрузки исходных данных, вычислений и текстового или графического вывода полученных результатов. Такая последовательность может быть сформирована пользователем как с помощью командной строки (интерактивный режим), так и из текстового файла (пакетный режим), а текстовые или графические результаты вычислений могут быть выведены на экран и/или записаны в соответствующие файлы.

Для пользователя, привычного к графическому интерфейсу, подобный подход может показаться неудобным и устаревшим, но, к счастью, это лишь широко распространенное заблуждение. После отработки основных навыков эффективность обработки данных с использованием клавиатуры и интерфейса командной строки оказываются не ниже, а выше, чем с помощью мыши и графического интерфейса. Одна из причин состоит в том, что вынести в меню и на пиктограммы сотни функций, применяемых в статистическом анализе крайне затруднительно, если вообще возможно, а командная строка R принимает любую комбинацию функций, корректную с точки зрения интерпретатора [1].

R предоставляет широкие возможности, которые могут быть продемонстрированы на примере анализа настроений пользователей социальной сети Твиттер.

Поздно ночью 26-го февраля среди пользователей начала распространяться информация о трагической гибели известного британского актера Роуэна Аткинсона. Этот факт не был подтвержден, но количество сообщений увеличивалось, и известие о смерти актера, сыгравшего мистера Бина, довольно быстро стало трендом Твиттера. Сообщения о кончине актера содержали фразу R.I.P. Rowan Atkinson. Используя возможности R можно проанализировать рассматриваемое событие и визуализировать этапы распространения информации.

Анализируемые данные представляют собой текстовый массив, содержащий информацию об отправителе, дате, времени и текст сообщения. Данные получены следующим образом:

```
library(twitteR)
tweets = searchTwitter("R.I.P. Rowan Atkinson", n=1500)
data = twListToDF(tweets).
```

Используя полученные данные, представляется возможным визуализировать процесс создания информационного потока и проследить время, когда пользователи проявляли наибольшую активность.

Рисунок 1 отображает генерируемое пользователями количество сообщений по дням. По представленному графику можно проследить появление первого сообщения, рост количества сообщений и достижение его пика в первой половине дня. Далее происходит постепенный спад и угасание интереса пользователей к этой теме.

График построен следующим образом:

```
library(ggplot2)
c <- ggplot(data, aes(created))
```

c + geom_bar()

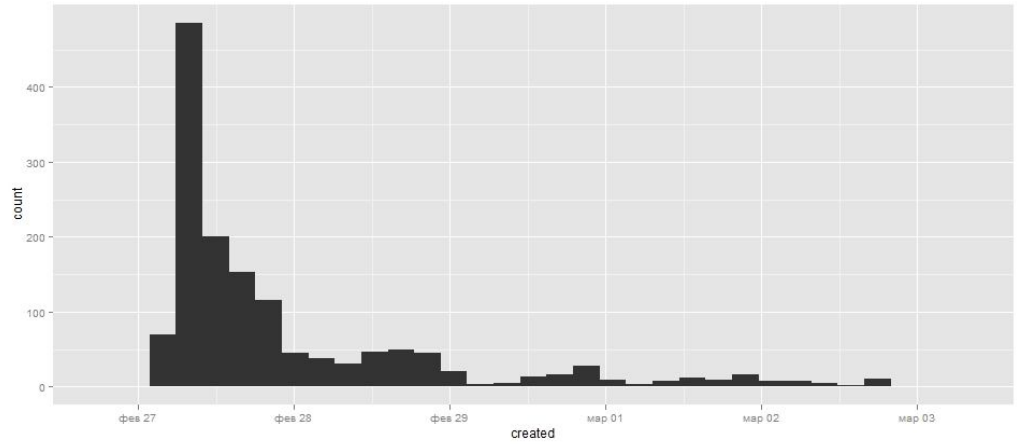

Рисунок 1. Количество сообщений в сутки

На рисунке 2 представлено распределение отправленных сообщений по часам. По графику видно, что первые сообщения начали появляться после 20 часов 26-го февраля, а наибольшее количество сообщений приходится на утро 27-го февраля.

График построен следующим образом:

```
library(ggplot2)
data$month=sapply(data$created, function(x) {p=as.POSIXlt(x);p$mon})
data$hour=sapply(data$created, function(x) {p=as.POSIXlt(x);p$hour})
data$wday=sapply(data$created, function(x) {p=as.POSIXlt(x);p$wday})
ggplot(data)+geom_jitter(aes(x=wday,y=hour))
```

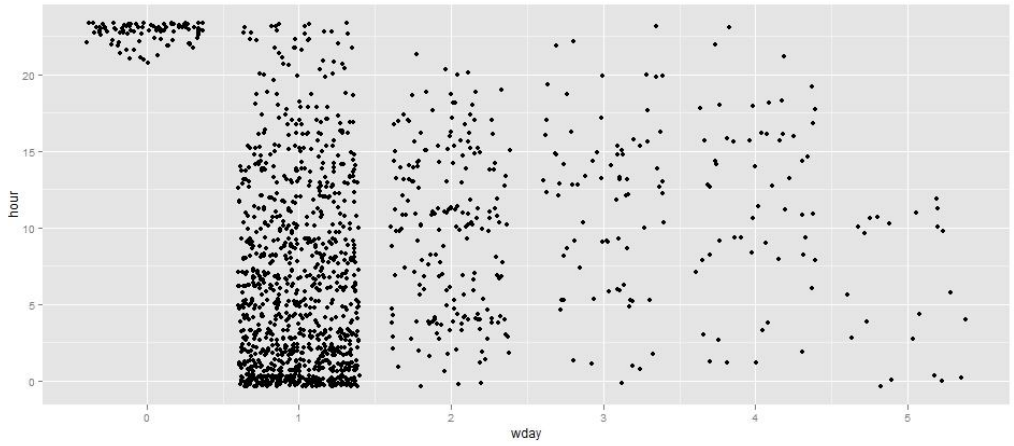

Рисунок 2. Распределение отправленных сообщений по часам

На рисунке 3 представлено облако из наиболее часто встречающихся слов в сообщениях пользователей. График построен следующим образом:

```
library("tm")
text = Corpus(DataframeSource(data.frame(data[1])))
text = tm_map(text, removePunctuation)
text = tm_map(text, tolower)
tdm = TermDocumentMatrix(text)
m = as.matrix(tdm)
v = sort(rowSums(m),decreasing=TRUE)
library("wordcloud")
```

wordcloud(names(v), v^0.3, scale=c(5,0.5),random.order=F, colors="black")

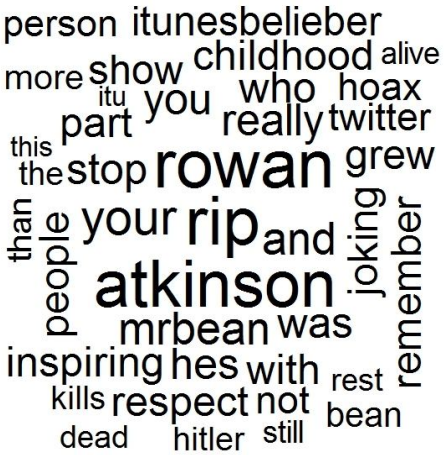

Рисунок 3. Облако наиболее встречающихся в сообщениях слов

Приведенный пример показывает, насколько широки возможности R. Уже на такой небольшой выборке данных можно проанализировать социальные аспекты распространения информации и выделить некоторые закономерности, используя при этом созданные в R графики.

В настоящее время реализации R существуют для трех наиболее распространенных семейств операционных систем: GNU/Linux, Apple Mac OS X и Microsoft Windows. В распределенных хранилищах системы CRAN по состоянию на конец сентября 2010 года были доступны для свободной загрузки 2548 пакетов расширения, ориентированных на специфические задачи обработки данных, возникающие в эконометрике и финансовом анализе, генетике и молекулярной биологии, экологии и геологии, медицине и фармацевтике и многих других прикладных областях. Значительная часть европейских и американских университетов в последние годы активно переходят к использованию R в учебной и научно-исследовательской деятельности вместо дорогостоящих коммерческих разработок.

Литература

1. Статистический анализ данных в системе R. Учебное пособие /А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г- Воронеж: ВГАУ, 2010. - 124с.