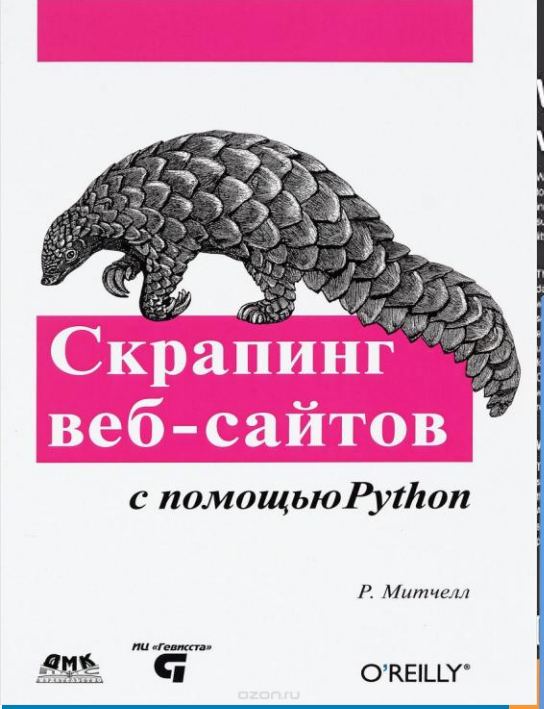
СПбГУТ им. проф. М.А.Бонч-Бруевича

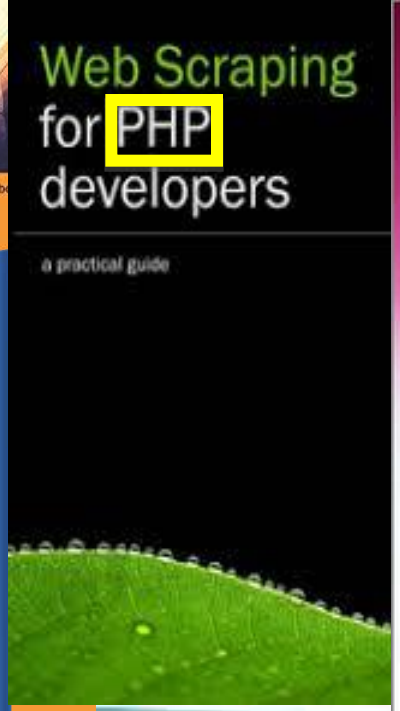
Web - скрапинг

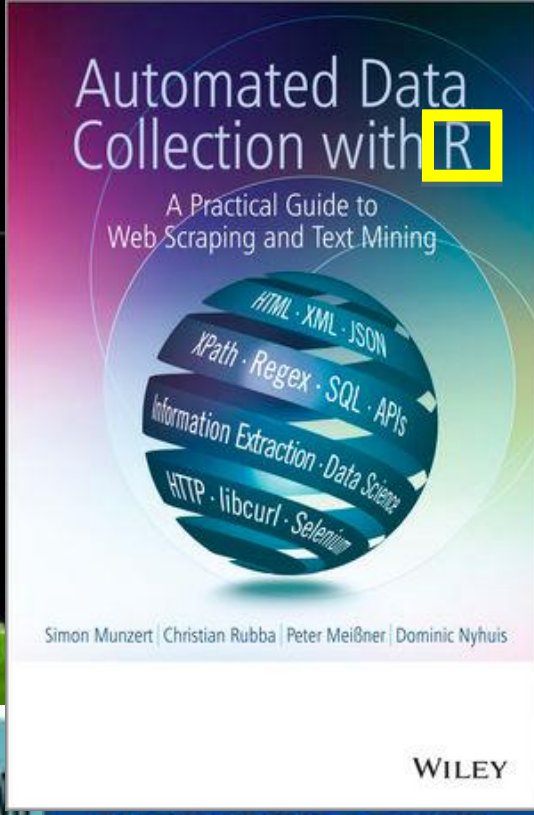
Доцент Филиппов Ф.В.

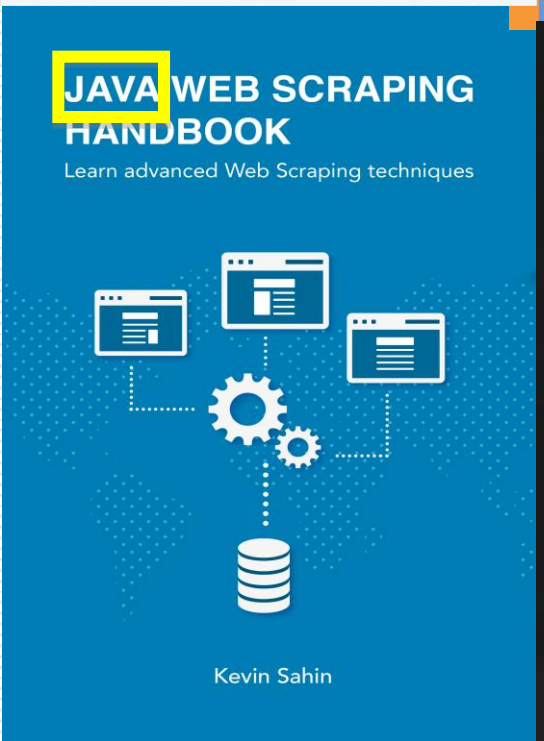
9000096@mail.ru


Скрапинг веб-сайтов
с помощью Python
Р. Митчелл
O'REILLY


Copyrighted Material
Web Scraping with Python
THE ULTIMATE GUIDE TO WEB SCRAPING
BY HARTLEY BRODY


Web Scraping for PHP developers
a practical guide


Automated Data Collection with R
A Practical Guide to Web Scraping and Text Mining
HTML · XML · JSON
XPath · Regex · SQL · APIs
Information Extraction · Data Science
HTTP · libcurl · Selenium
Simon Munzert | Christian Rubba | Peter Meißner | Dominic Nyhuis
WILEY


JAVA WEB SCRAPING HANDBOOK
Learn advanced Web Scraping techniques
Kevin Sahin


INSTANT
Short | Fast | Focused
PHP Web Scraping
Jacob Ward
PACKT


WEB SCRAPING
INDUSTRY TRENDS
FOR 2017


WEB SCRAPING WITH PYTHON IN A DAY
THE ULTIMATE CRASH COURSE TO LEARNING THE BASICS OF WEB SCRAPING WITH PYTHON IN NO TIME

Что такое *web*-скрапинг?

Веб-скрапинг (*Web Scraping*) - совокупность методов получения интересующего контента с небольшими затратами.

Иногда данные методы носят название «парсинга контента» или «парсинга сайтов».

Методы состоят в получении веб-страницы и обработке её содержимого по специальному алгоритму, осуществляющему сбор необходимой информации.

Web Scraping

Extract data from any website

Проблемы (задачи) *web*-скрапинга

- Проблема навигации: поиск и получение целевых страниц для извлечения данных.
- Проблема распознавания данных: распознавание участков, содержащих нужную информацию.
- Проблема поиска общей структуры данных: структурирование извлекаемой информации.
- Проблема сопоставления атрибутов полученных данных: обеспечение однородности информации.
- Проблема объединения данных: объединение информации, полученной из разных источников.

Вопросы для изучения

1. Правовые и этические аспекты *web*-скрапинга
2. Технологии размещения, извлечения и сохранения *web*-данных
3. Методики *web*-скрапинга

1

Правовые и этические аспекты *web*-скрапинга

Вредоносный *web*-скрапинг

- Несанкционированый доступе к защищённым специальными мерами безопасности ресурсам с целью хищения информации.
- Публикация без разрешения автора собранного материала, защищённого законами об авторском праве.
- Сбор персональной информации.
- Длительные и массированные запросы к веб-ресурсу, которые могут затруднить его работу и сделать сайт временно недоступным, вызвав ситуацию «отказ в обслуживании» (DDOS-атака).
- Нарушение установленным ресурсом правил, обычно размещённых на главной странице сайта и в файле ***robots.txt***.

Файл *robots.txt*

Файл *robots.txt* появился в 1994 году в результате дискуссий о том, как защитить от поисковых роботов информацию, которую владельцы ресурсов хотели бы сохранить доступной, но в то же время сделать непубличной. По результатам дискуссии был принят Стандарт исключений для роботов (*Robots Exclusion Standard*).

Правила Facebook являются одними из наиболее строгих: https://www.facebook.com/robots.txt. В находящейся здесь ссылке содержится свод правил, одно из которых прямо запрещает краулинг и скрапинг этого ресурса целиком всем, кроме перечня агентов, для которых также прописаны запретные места.

Файл *robots.txt*

Синтаксис *robots.txt* предусматривает возможность указания роботам и скраперам частоты обращения к страницам ресурса, в секундах:

User-agent: YandexNews
Crawl-delay: 2

При нарушении последнего правила, *IP*-адрес, с которого выполняются запросы будет автоматически заблокирован – постоянно или на некоторый установленный администратором срок.

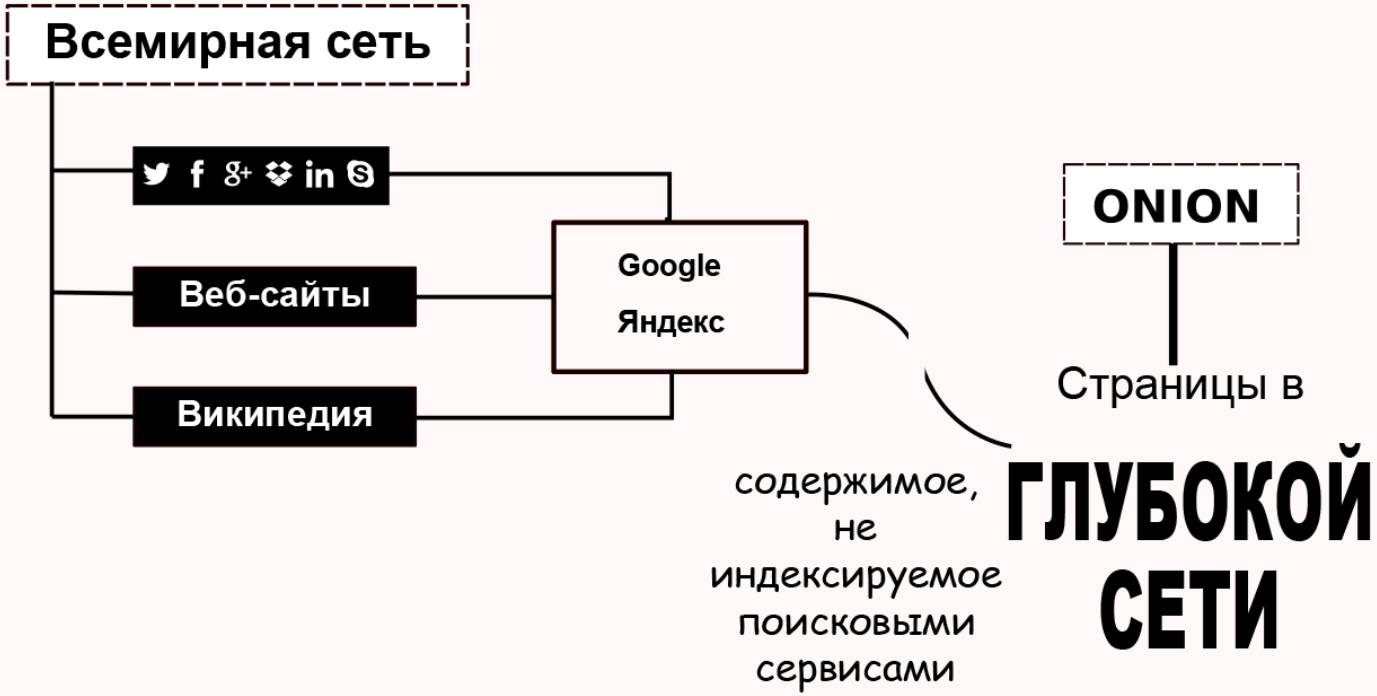
Использование метатегов

Для указания средствам автоматизированного сбора информации правил доступа к конкретной странице в её *HTML*-код предлагается помещать специальные теги:

```
<html>
<head>
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
</head>
```

Запрещение агентам поисковых систем обходить и индексировать веб-страницы наряду с использованием специальных ловушек лежит в основе разделения пространства интернета на видимый (*Clear Internet*) и «глубокий» (*Deep Internet*) – содержимое относящихся к последней категории страниц хоть и находится в открытом доступе, но его поиск через Google, Yandex и другие системы не приносит результата.

Deep Internet



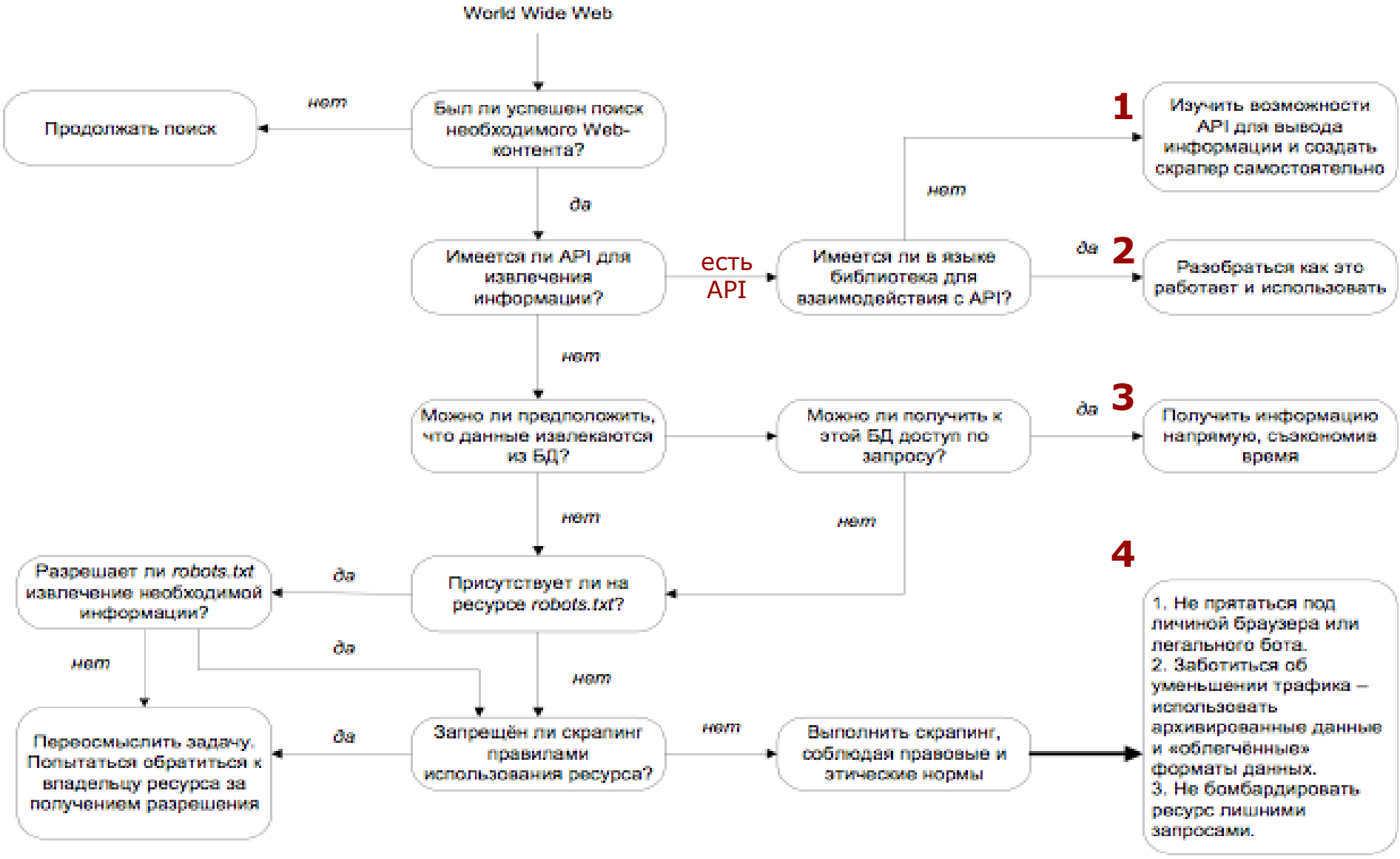
https://ru.wikipedia.org/wiki/Глубокая_паутина

Когда используется *web*-скрапинг?

Для получения информации из веб-ресурсов не всегда требуется применять метод скрапинга. Если существуют другие, более оптимальные средства, то использовать следует именно их. Если же скрапинга не избежать, то разрабатываемая программа-скрапер должна отвечать следующим требованиям в порядке убывания их значимости:

- Следовать правилам использования ресурсов и сложившимся нормам поведения в глобальной сети.
- Не перегружать информационный ресурс запросами и извлекать из него только действительно необходимые данные.
- Быть эффективной.

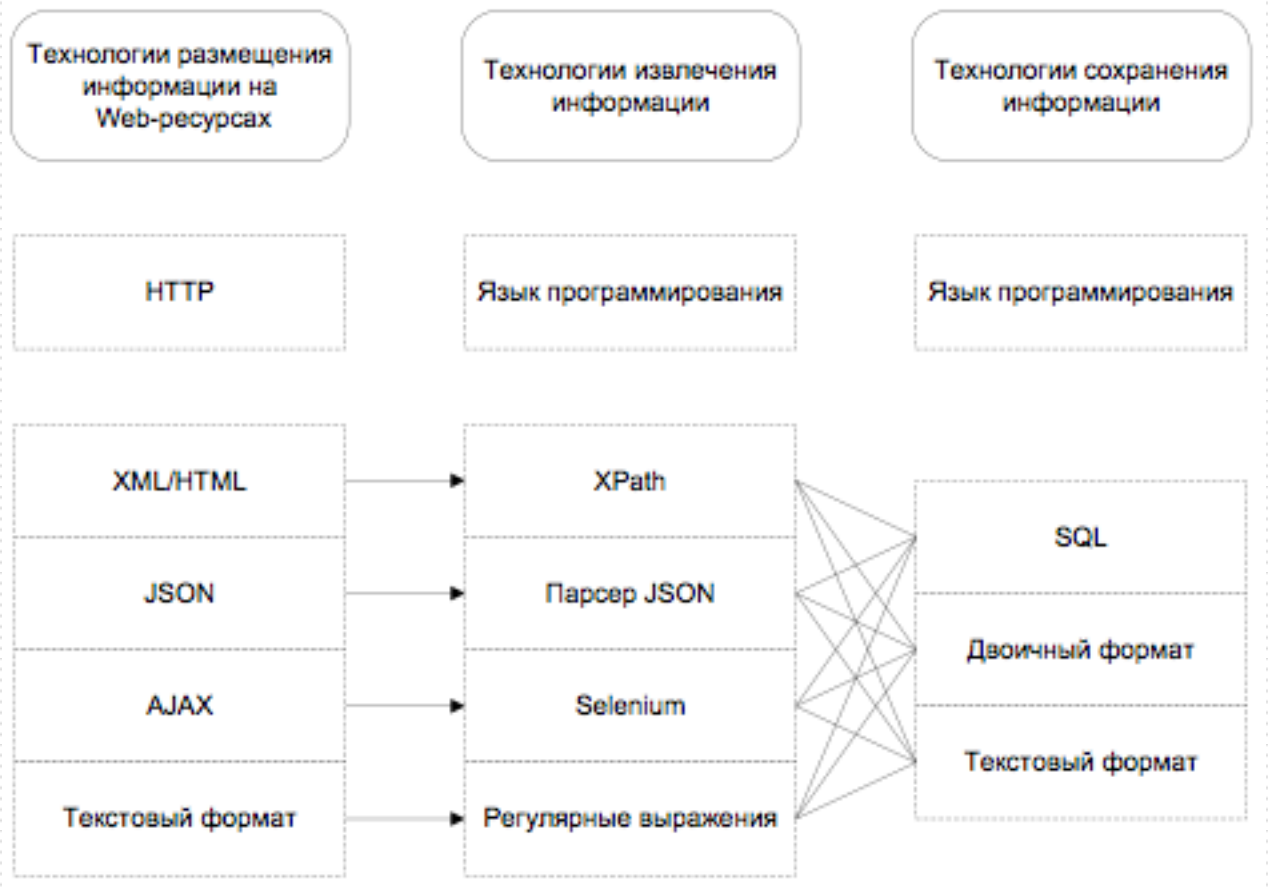
Этичная методика получения информации из веб-ресурсов



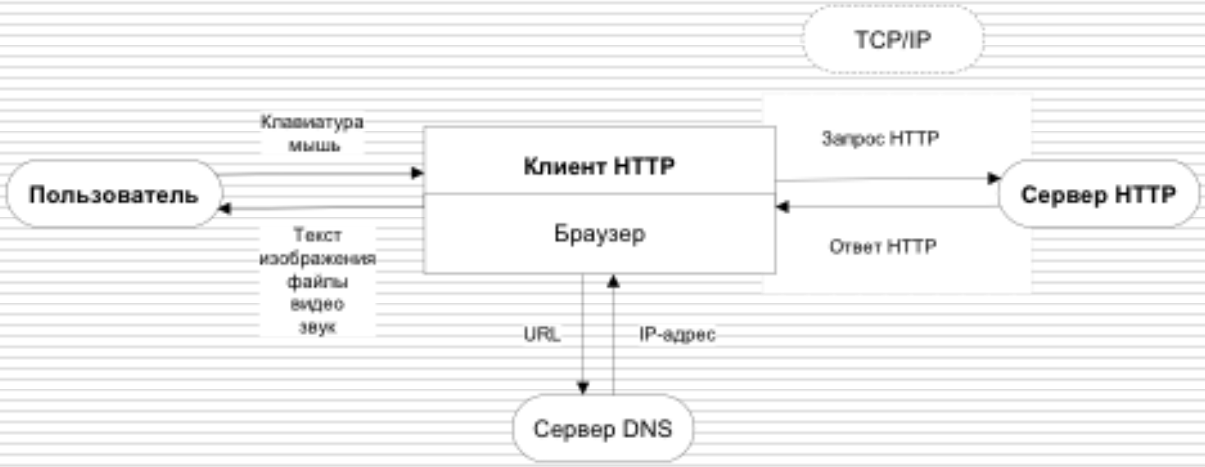
2

Технологии размещения, извлечения и сохранения *web*-данных

Классификация технологий



HTTP



1xx – состояние обработки запроса.
2xx – успешное выполнение запроса.
3xx – необходима переадресация.
4xx – коды ошибок на стороне клиента.
5xx – коды ошибок на стороне сервера.

URL (Uniform Resource Locator)

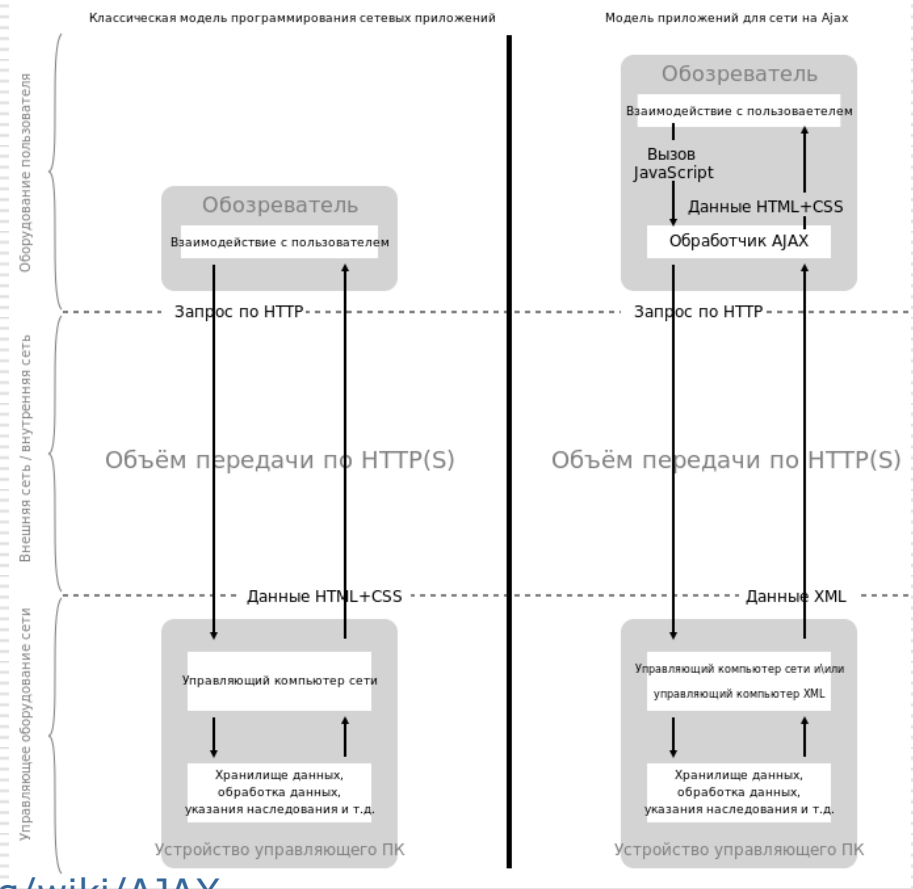
scheme://hostname:port/path?querystring#fragment

scheme – протокол обмена
hostname - уникальный идентификатор сервера
port – «дверь» сервера (своя для каждого протокола)
path - путь доступа к файловому ресурсу
querystring - уточнение расположения запрашиваемой информации
fragment - указывает на конкретную часть документа

HTML

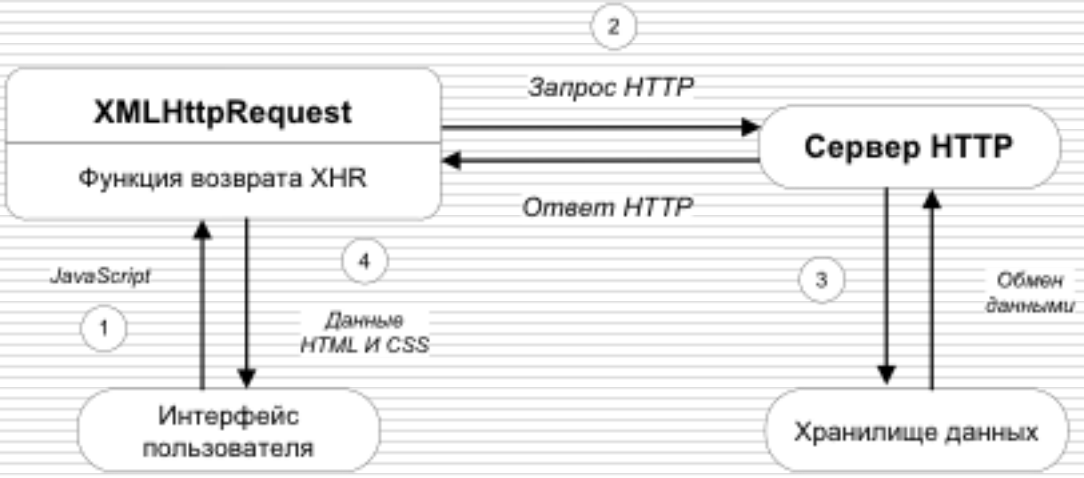
- Обеспечение полной совместимости версий *HTML*-документов сверху вниз.
- *HTML*-документ представляет собой обычный текстовый документ, который можно открыть в любом редакторе.
- Имеется возможность использовать в исходном документе гиперссылки, которые образуют цепочку связей с другими документами, не обязательно гипертекстовыми.
- Для расширения функциональности содержит специальный тег <*script*>, который служит контейнером для скриптов, позволяющих использовать возможности языков программирования.
- Полностью структурирован в соответствии с *DOM*.

AJAX (Asynchronous JavaScript and XML)



https://ru.wikipedia.org/wiki/AJAX

XHR (XMLHttpRequest)



API, доступное в скриптовых языках браузеров, таких как JavaScript. Использует запросы HTTP или HTTPS напрямую к веб-серверу и загружает данные ответа сервера напрямую в вызывающий скрипт. Информация может передаваться в любом текстовом формате, например, в XML, HTML или JSON. Позволяет осуществлять HTTP-запросы к серверу без перезагрузки страницы.

Структурированность данных

- Полностью структурированные. Такие данные представлены в виде таблиц, которые легко извлекаемы и не требуют специальных преобразований для сохранения в табличном виде.
- Хорошо структурированные. Информация на страницах использует для размещения шаблоны. Путь к необходимым данным задаётся при помощи *CSS*-селекторов или выражений *XPath*.
- Плохо структурированные. Однотипные данные размещены на странице с использованием разных элементов разметки.
- Неструктурированные. Структура данных полностью отсутствует. Информация с таких страниц извлекается построчно.

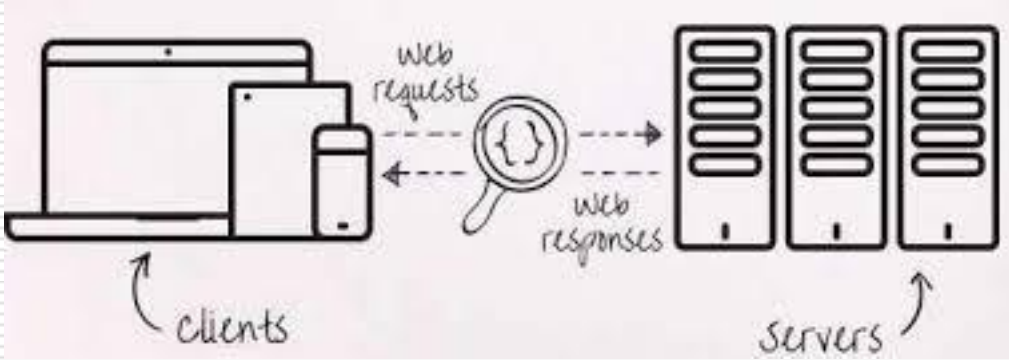
3

Методики *web*-скрапинга

Правила сбора информации

- Точно определить, какого рода информация требуется.
- Определить источники информации в глобальной сети наиболее точно соответствующие поставленному вопросу. Например, твиты содержат личные мнения людей практически обо всём, а анализ информации о продажах интернет-магазинов покажет их действительные предпочтения.
- Сформировать представление о процессе происхождения информации на её потенциальном источнике – когда она появилась, когда и кем была загружена.
- Соблюсти баланс между преимуществами и недостатками извлечения информации – общедоступность и легальность, цена извлечения, совместимость с другими источниками для её оценки и т.п.

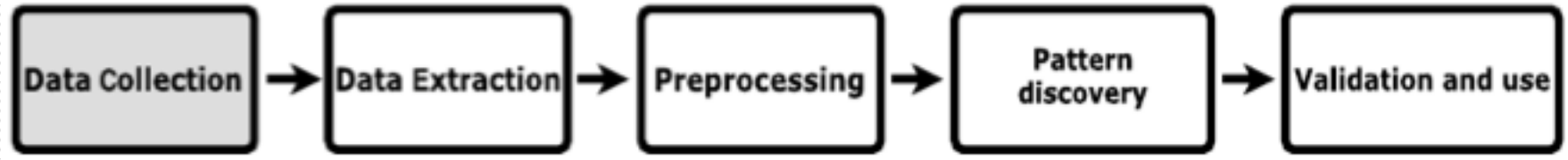
Разработка Web-скрапера



- Для перехода по страницам ресурса необходимо обеспечить автоматическое формирование URL-адресов.
- Для извлечения из контента ресурса требуемой информации необходимо обеспечить ограничение поиска.
- Разработанный Web-скрапер должен обладать достаточной функциональностью для извлечения информации из аналогичных ресурсов.

Пакеты языка R

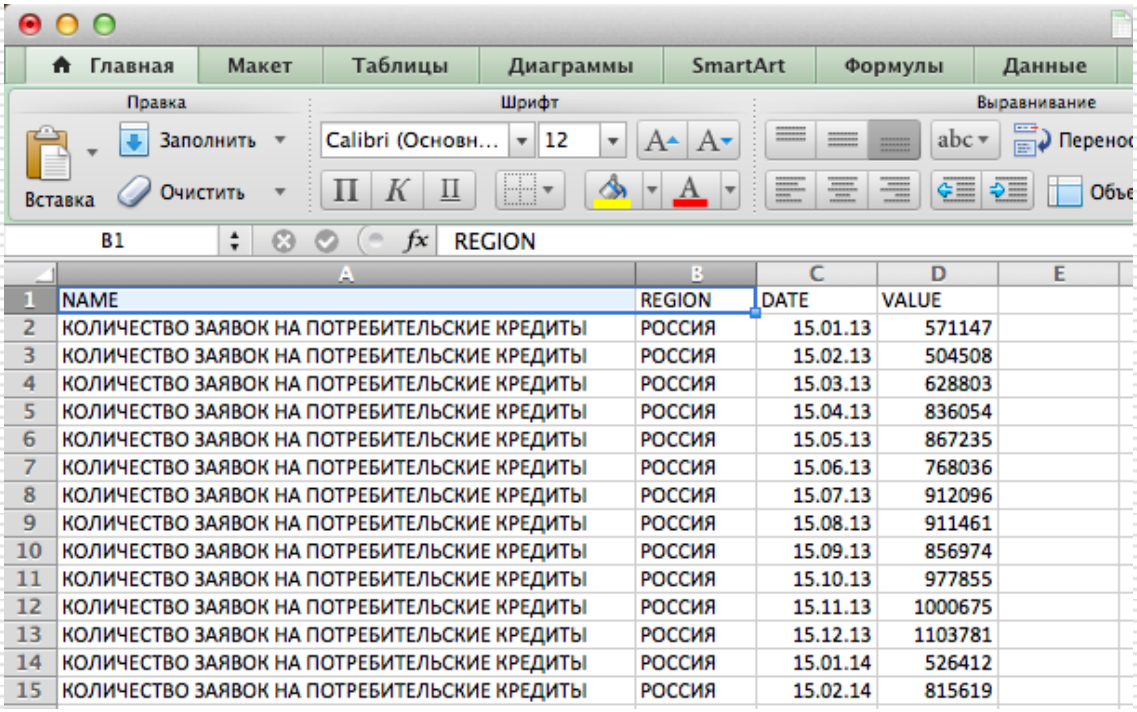
Package Name	Crawl	Retrieve	Parse	Description
scrapeR	No	Yes	Yes	From a given vector of URLs, retrieves web pages and parses them to pull out information of interest using an XPath pattern.
tm.plugin.webmining	No	Yes	Yes	Follows links on web feed formats like XML and JSON, and extracts contents using boilerpipe method.
Rvest	No	Yes	Yes	Wraps around the *xml2* and *httr* packages so that they can easily to download and then manipulate HTML and XML.
RCrawler	Yes	Yes	Yes	Crawls web sites and extracts their content using various techniques.
Some basic web toolkits:				
XML, XML2	No	No	Yes	HTML / XML parsers
jsonlite, RJSONIO	No	No	Yes	JSON parser
RSelenium	No	No	Yes	Browser Automation
Selectr	No	No	Yes	Parses CSS3 Selectors and translates them to XPath 1.0
Httr, RCurl	No	Yes	No	Handles HTTP / HTTPS requests



RCrawler: An R package for parallel web crawling and scraping
Salim Khalil, Mohamed Fakir

Извлечение полностью структурированных (статических) данных

Копируем данные из http://www.sberbank.com/ru/analytics/opendata



	A	B	C	D	E
1	NAME	REGION	DATE	VALUE	
2	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.01.13	571147	
3	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.02.13	504508	
4	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.03.13	628803	
5	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.04.13	836054	
6	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.05.13	867235	
7	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.06.13	768036	
8	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.07.13	912096	
9	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.08.13	911461	
10	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.09.13	856974	
11	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.10.13	977855	
12	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.11.13	1000675	
13	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.12.13	1103781	
14	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.01.14	526412	
15	КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ	РОССИЯ	15.02.14	815619	

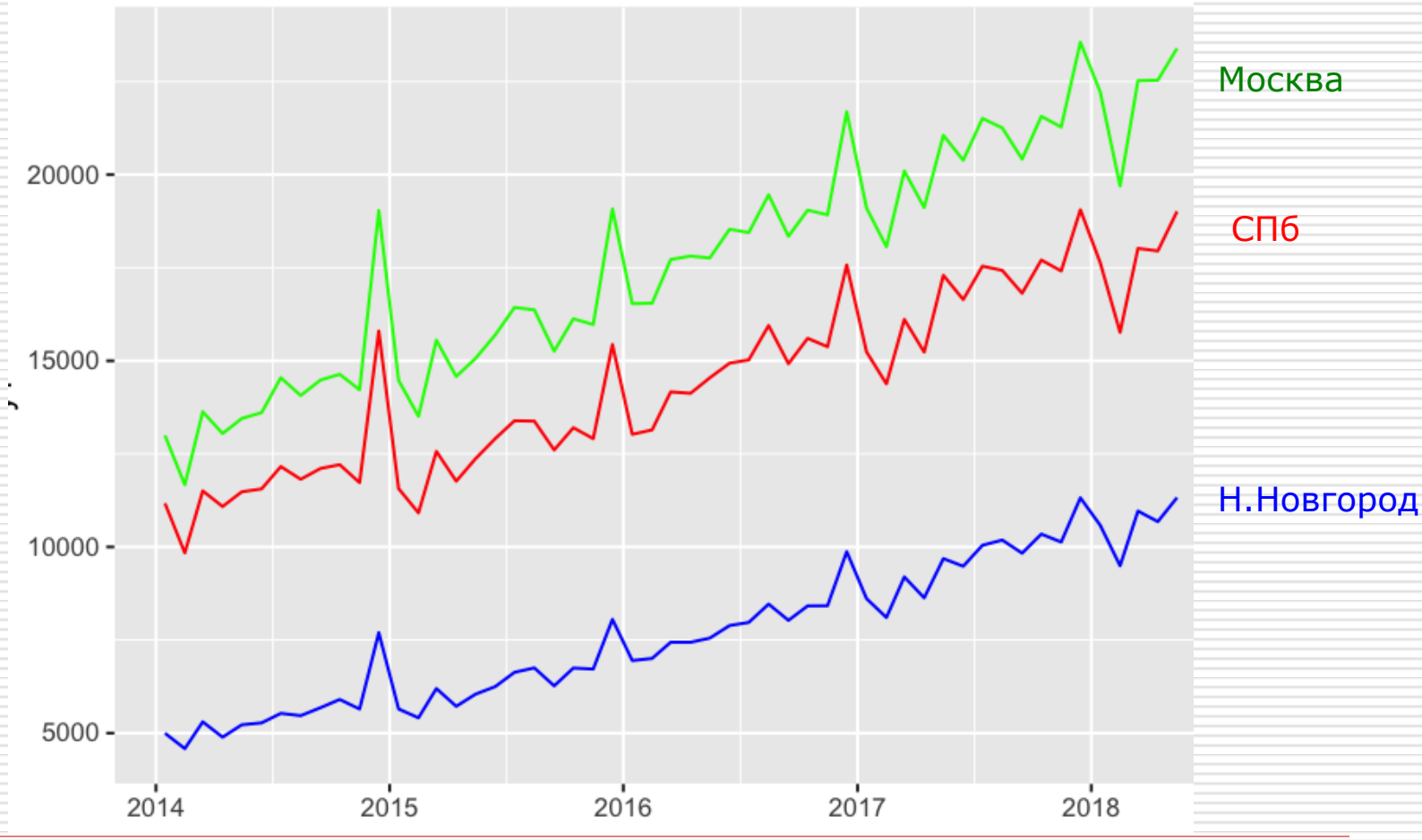
Извлечение полностью структурированных (статических) данных

```
1  library("readxl")
2  sd <- read_excel("/Users/mac/Desktop/LABS/sbData.xlsx")
3  wh01 <- unique(sd$NAME) # параметры (15)
4  #[1] "КОЛИЧЕСТВО ЗАЯВОК НА ПОТРЕБИТЕЛЬСКИЕ КРЕДИТЫ"   "СРЕДНЯЯ СУММА ЗАЯВКИ НА ПОТРЕБИТЕЛЬСКИЙ КРЕДИТ"
5  #[3] "КОЛИЧЕСТВО ЗАЯВОК НА ИПОТЕЧНЫЕ КРЕДИТЫ"        "СРЕДНЯЯ СУММА ЗАЯВКИ НА ИПОТЕЧНЫЙ КРЕДИТ"
6  #[5] "КОЛИЧЕСТВО НОВЫХ ДЕПОЗИТОВ"                    "СРЕДНЯЯ СУММА НОВОГО ДЕПОЗИТА"
7  #[7] "СРЕДНЯЯ ЗАРПЛАТА"                              "СРЕДНЯЯ ПЕНСИЯ"
8  #[9] "В СРЕДНЕМ РУБ. НА ТЕКУЩЕМ СЧЕТЕ НА ЧЕЛОВЕКА"   "В СРЕДНЕМ ДЕПОЗИТОВ В РУБ. НА ЧЕЛОВЕКА"
9  #[11] "СРЕДНИЕ РАСХОДЫ ПО КАРТАМ"                    "СРЕДНИЙ ЧЕК В ФОРМАТЕ ФАСТФУД"
10 #[13] "СРЕДНИЙ ЧЕК В ФОРМАТЕ РЕСТОРАН"               "СРЕДНИЕ ТРАТЫ В РЕСТОРАНЕ ФАСТФУД"
11 #[15] "СРЕДНИЕ ТРАТЫ В РЕСТОРАНЕ"
12 wh02 <- unique(sd["REGION"]) # регионы (84)
```

Извлечение полностью структурированных (статических) данных

```r
sd <- read_excel("/Users/mac/Desktop/LABS/sbData.xlsx")
wh01 <- unique(sd$NAME) # параметры (15)
wh02 <- unique(sd["REGION"]) # регионы (84)
wh03 <- unique(sd[3]) # даты (65)

spb1 <- sd$REGION[sd$REGION == "САНКТ-ПЕТЕРБУРГ"]
y1 <- sd$VALUE[sd$REGION == "САНКТ-ПЕТЕРБУРГ" & sd$NAME == "СРЕДНИЕ РАСХОДЫ ПО КАРТАМ"]
y2 <- sd$VALUE[sd$REGION == "МОСКВА" & sd$NAME == "СРЕДНИЕ РАСХОДЫ ПО КАРТАМ"]
x <- sd$DATE[sd$REGION == "САНКТ-ПЕТЕРБУРГ" & sd$NAME == "СРЕДНИЕ РАСХОДЫ ПО КАРТАМ"]
df <- data.frame(x,y1,y2)

y3 <-sd$VALUE[sd$REGION == "НИЖЕГОРОДСКАЯ ОБЛАСТЬ" & sd$NAME == "СРЕДНИЕ РАСХОДЫ ПО КАРТАМ"]

# https://r-datascience.ru/ggplot2_guide/
library("ggplot2")
ggplot(df, aes(x)) +
  geom_line(aes(y=y1), colour="red") +
  geom_line(aes(y=y2), colour="green") +
  geom_line(aes(y=y3), colour="blue")
```

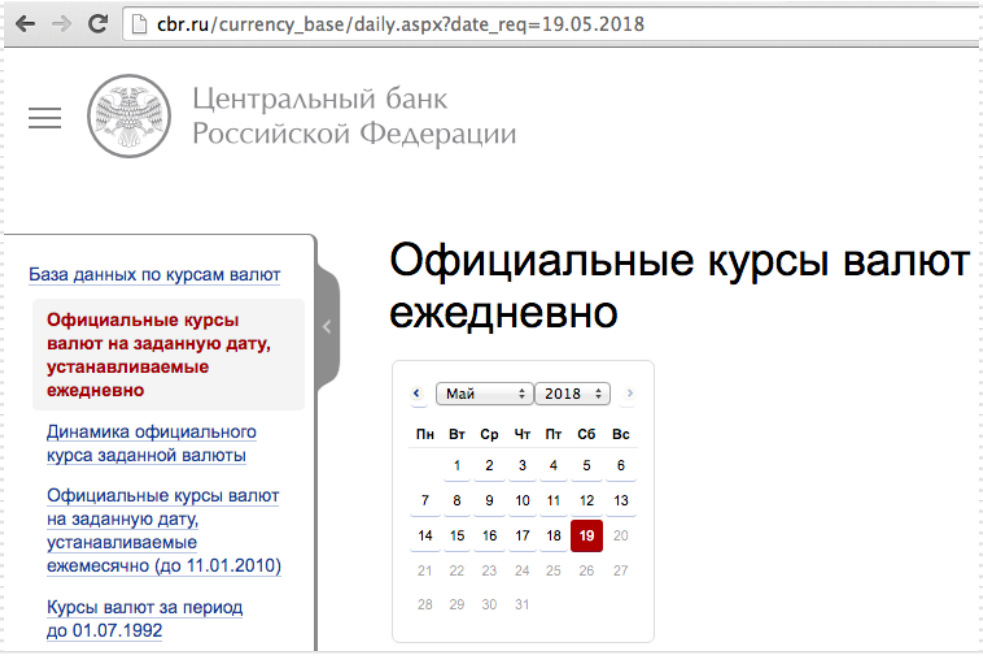
Средние расходы по картам Сбербанка



Извлечение полностью структурированных (динамических) данных

http://cbr.ru/currency_base/daily.aspx

```
library("rvest")
library("ggplot2”)
```

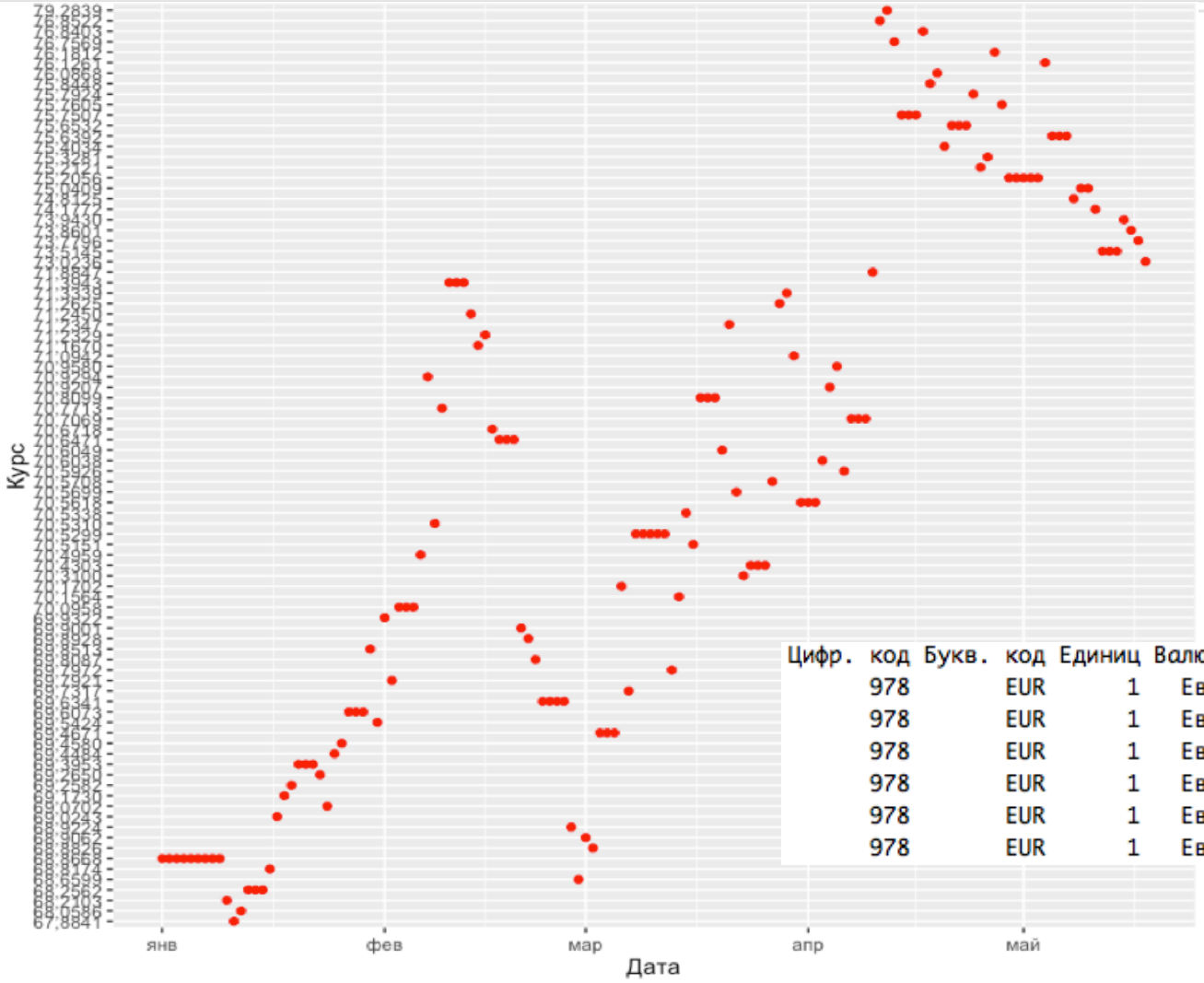


Параметры функции:

```
Rates <- function(BaseURL, BegDate, EndDate, Curr)
```

```
library("rvest")
library("ggplot2")

Rates <- function(BaseURL, BegDate, EndDate, Curr) {
  bd <- as.Date(BegDate)
  ed <- as.Date(EndDate)
  vDate <- seq.Date(bd, ed, 1)
  Currency <- toupper(Curr)
  fCourse = NULL

  len <- length(vDate)
  for (dd in 1:len)
  {
    locURL <- paste0(BaseURL, vDate[dd])
    docSource <- read_html(locURL)
    table <- html_table(docSource)
    tab <- table[[1]]
    fCourse <- rbind(fCourse,cbind(subset(tab, tab[2]==Currency),
dDate=vDate[dd]))
    Sys.sleep(3)
  }
  fCourse
}
```

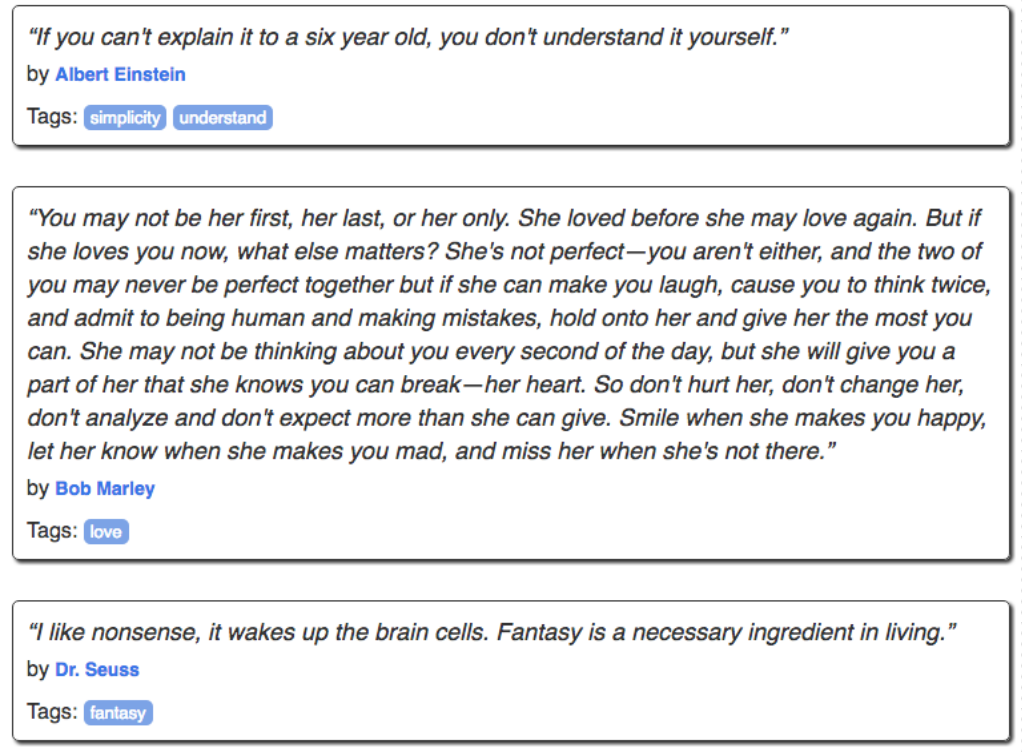
```
z <- Rates("http://cbr.ru/currency_base/daily.aspx?date_req=", "2018-01-01", "2018-05-18", "eur")

ggplot(data=z, aes(x=z$"dDate",y=z$"Курс")) + geom_point(color = "red") + labs (x="Дата",y="Курс")
```



Цифр. код	Букв. код	Единиц	Валюта	Курс	dDate
978	EUR	1	Евро	72,5806	2018-06-01
978	EUR	1	Евро	72,7183	2018-06-02
978	EUR	1	Евро	72,7183	2018-06-03
978	EUR	1	Евро	72,7183	2018-06-04
978	EUR	1	Евро	72,4384	2018-06-05
978	EUR	1	Евро	72,5130	2018-06-06

Извлечение структурированных данных с динамических *web*-страниц

http://spidyquotes.herokuapp.com/scroll



Анализ структуры подгружаемого ресурса

http://spidyquotes.herokuapp.com/api/quotes?page=3


```
{
  "has_next": true,
  "page": 3,
  "quotes": [
    {
      "author": {
        "goodreads_link": "/author/show/4026.Pablo_Neruda",
        "name": "Pablo Neruda",
        "slug": "Pablo-Neruda"
      },
      "tags": [
        "love",
        "poetry"
      ],
      "text": "\u201cI love you without knowing how, or when, or from where. I
of loving but this, in which there is no I or you, so intimate that your hand
    },
```

Построение скрапера

```
library(jsonlite)
BaseUrl <- "http://spidyquotes.herokuapp.com/api/quotes?page="
CurPage <- 0
has_next <- TRUE
DF = NULL
pages <- list()

while (has_next == TRUE)
{
  url <- paste0(BaseUrl, CurPage + 1)
  JSpart <- fromJSON(url)
  has_next <- JSpart$has_next
  CurPage <- JSpart$page
  pages[[CurPage]] <- JSpart$quotes
  Sys.sleep(1)
}
FD <- as.data.frame(pages)
```

Извлеченная информация

	author.goodreads_link	author.name	author.slug	tags	text
1	/author/show/9810.Albert_Einstein	Albert Einstein	Albert-Einstein	c("change", "deep-thoughts", "thinking", "world")	"The world as we have created it is a process of our thi...
2	/author/show/1077326.J_K_Rowling	J.K. Rowling	J-K-Rowling	c("abilities", "choices")	"It is our choices, Harry, that show what we truly are, f...
3	/author/show/9810.Albert_Einstein	Albert Einstein	Albert-Einstein	c("inspirational", "life", "live", "miracle", "miracles")	"There are only two ways to live your life. One is as tho...
4	/author/show/1265.Jane_Austen	Jane Austen	Jane-Austen	c("aliteracy", "books", "classic", "humor")	"The person, be it gentleman or lady, who has not plea...
5	/author/show/82952.Marilyn_Monroe	Marilyn Monroe	Marilyn-Monroe	c("be-yourself", "inspirational")	"Imperfection is beauty, madness is genius and it's bet...
6	/author/show/9810.Albert_Einstein	Albert Einstein	Albert-Einstein	c("adulthood", "success", "value")	"Try not to become a man of success. Rather become a...
7	/author/show/7617.Andr_Gide	André Gide	Andre-Gide	c("life", "love")	"It is better to be hated for what you are than to be lov...
8	/author/show/3091287.Thomas_A_Edison	Thomas A. Edison	Thomas-A-Edison	c("edison", "failure", "inspirational", "paraphrased")	"I have not failed. I've just found 10,000 ways that won...
9	/author/show/44566.Eleanor_Roosevelt	Eleanor Roosevelt	Eleanor-Roosevelt	misattributed-eleanor-roosevelt	"A woman is like a tea bag; you never know how stron...
10	/author/show/7103.Steve_Martin	Steve Martin	Steve-Martin	c("humor", "obvious", "simile")	"A day without sunshine is like, you know, night."

author.1.goodreads_link	author.1.name	author.1.slug	tags.1	text.1
/author/show/82952.Marilyn_Monroe	Marilyn Monroe	Marilyn-Monroe	c("friends", "heartbreak", "inspirational", "life", "love", "...	"This life is what you make it. No matter what, you're g...
/author/show/1077326.J_K_Rowling	J.K. Rowling	J-K-Rowling	c("courage", "friends")	"It takes a great deal of bravery to stand up to our ene...
/author/show/9810.Albert_Einstein	Albert Einstein	Albert-Einstein	c("simplicity", "understand")	"If you can't explain it to a six year old, you don't unde...
/author/show/25241.Bob_Marley	Bob Marley	Bob-Marley	love	"You may not be her first, her last, or her only. She lov...
/author/show/61105.Dr_Seuss	Dr. Seuss	Dr-Seuss	fantasy	"I like nonsense, it wakes up the brain cells. Fantasy is...
/author/show/4.Douglas_Adams	Douglas Adams	Douglas-Adams	c("life", "navigation")	"I may not have gone where I intended to go, but I thin...
/author/show/1049.Elie_Wiesel	Elie Wiesel	Elie-Wiesel	c("activism", "apathy", "hate", "indifference", "inspiratio...	"The opposite of love is not hate, it's indifference. The ...
/author/show/1938.Friedrich_Nietzsche	Friedrich Nietzsche	Friedrich-Nietzsche	c("friendship", "lack-of-friendship", "lack-of-love", "love...	"It is not a lack of love, but a lack of friendship that m...
/author/show/1244.Mark_Twain	Mark Twain	Mark-Twain	c("books", "contentment", "friends", "friendship", "life")	"Good friends, good books, and a sleepy conscience: t...
/author/show/276029.Allen_Saunders	Allen Saunders	Allen-Saunders	c("fate", "life", "misattributed-john-lennon", "planning", ...	"Life is what happens to us while we are making other ...

Задания по скрапингу

1. По материалам сладов 31-33 построить график изменения курса доллара за 30 последних дней. Использовать ресурс http://cbr.ru.

2. По материалам слайдов 35-38 построить фрейм, включающий два столбца – автор (author) и текст (text). Для всех авторов из ресурса http://spidyquotes.herokuapp.com/scroll.

Crawler (ползунок)

Поиско́вый ро́бот («веб-пау́к», «веб-краулер») — программа, предназначенная для перебора страниц Интернета с целью получения информации о них.

Порядок обхода страниц, частота визитов, а также критерии выделения значимой информации определяются алгоритмами **информационного поиска**. *(В большинстве случаев переход от одной страницы к другой осуществляется по ссылкам, содержащимся на первой и последующих страницах.)*

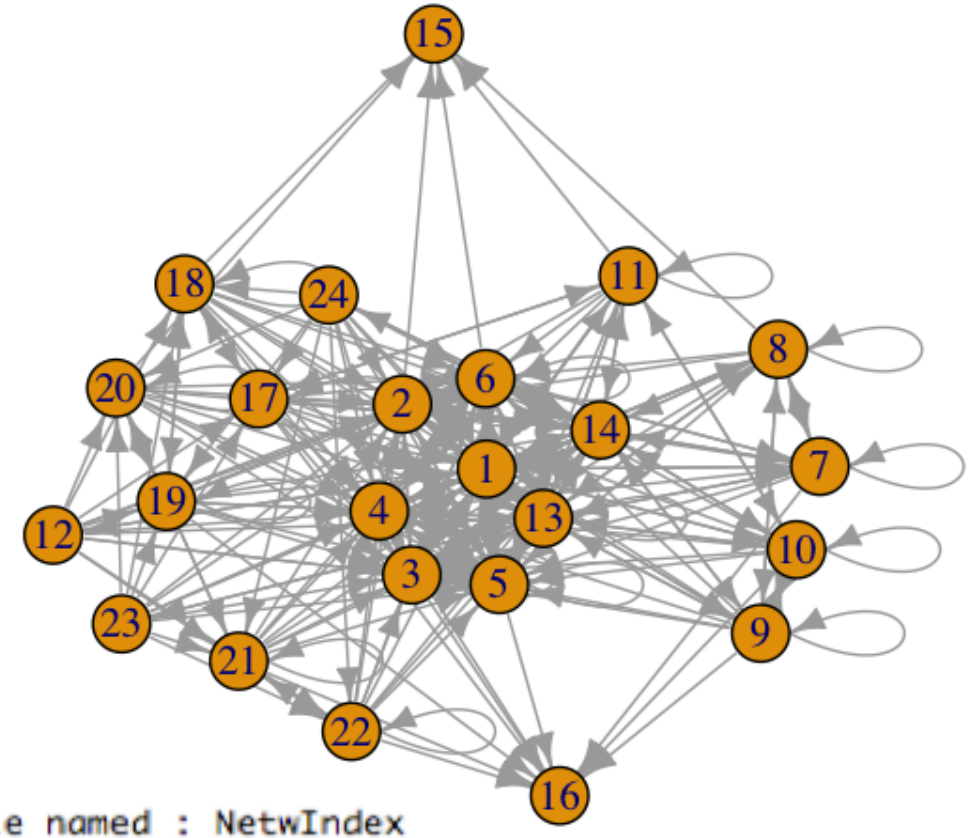
Информацио́нный по́иск (information retrieval) — процесс поиска неструктурированной информации, удовлетворяющей информационные потребности, и наука об этом поиске.

Пакет *RCrawler*

```
library(Rcrawler)
Rcrawler(Website = "http://glofile.com/" , NetworkData = TRUE)
# Проползет весь сайт и сформирует сеть из вершин внутренних ссылок.
library(igraph)
network<-graph.data.frame(NetwEdges, directed=T)
plot(network)
```

```
> Rcrawler(Website = "http://glofile.com/" , NetworkData = TRUE)
In process : 1..
Progress: 7.69 %  :  1  parssed from  13  | Collected pages: 1  | Level: 1
In process : 2..3..4..
Progress: 12.50 %  :  2  parssed from  16  | Collected pages: 4  | Level: 1
In process : 5..6..7..
Progress: 31.25 %  :  5  parssed from  16  | Collected pages: 7  | Level: 1
In process : 8..9..10..
Progress: 50.00 %  :  8  parssed from  16  | Collected pages: 10  | Level: 1
In process : 11..12..13..
Progress: 47.83 %  :  11  parssed from  23  | Collected pages: 13  | Level: 2
In process : 14..15..16..
Progress: 58.33 %  :  14  parssed from  24  | Collected pages: 14  | Level: 2
In process : 17..18..19..
Progress: 70.83 %  :  17  parssed from  24  | Collected pages: 17  | Level: 2
In process : 20..21..22..
Progress: 83.33 %  :  20  parssed from  24  | Collected pages: 20  | Level: 2
In process : 23..24..
Progress: 95.83 %  :  23  parssed from  24  | Collected pages: 22  | Level: 3
+ Check INDEX dataframe variable to see crawling details
+ Collected web pages are stored in Project folder
+ Project folder name : glofile.com-311416
+ Project folder path : /Users/mac/Desktop/R_script/glofile.com-311416
+ Network nodes are stored in a variable named : NetwIndex
+ Network eadges are stored in a variable named : NetwEdges
```

```
library(igraph)
network<-graph.data.frame(NetwEdges, directed=T)
plot(network)
```




```
Network nodes are stored in a variable named : NetwIndex
Network eadges are stored in a variable named : NetwEdges
```

Резюме

- Веб-скрапинг (*Web Scraping*) - совокупность методов получения интересующего контента с небольшими затратами.
- Веб-скрапинг широко используемая технология поиска неструктурированной информации, удовлетворяющей информационны.