Экстренные сообщения Росгидромета

Поиск элементов на сайте

Установим и загрузим пакет rvest для веб-скрапинга.

```
install.packages("rvest")
```

```
library(rvest)
```

Установим и загрузим пакет stringr для обработки строк.

```
install.packages("stringr")
```

```
library(stringr)
```

Установим и загрузим пакет tm для текст майнинга.

```
install.packages("tm")
```

```
library(tm)
```

На сайте Росгидромета по ссылке https://www.meteorf.ru/product/emergency размещены экстренные сообщения о погодных явлениях в разных регионах России с указанием даты и времени сообщения.

На странице выводятся по 10 сообщений. Ссылка на первую страницу:

http://www.meteorf.ru/product/emergency/?PAGEN_1=1.

На вторую:

http://www.meteorf.ru/product/emergency/?PAGEN_1=2.

То есть для того чтобы перебрать все нужные страницы, достаточно менять последнее число в ссылке.

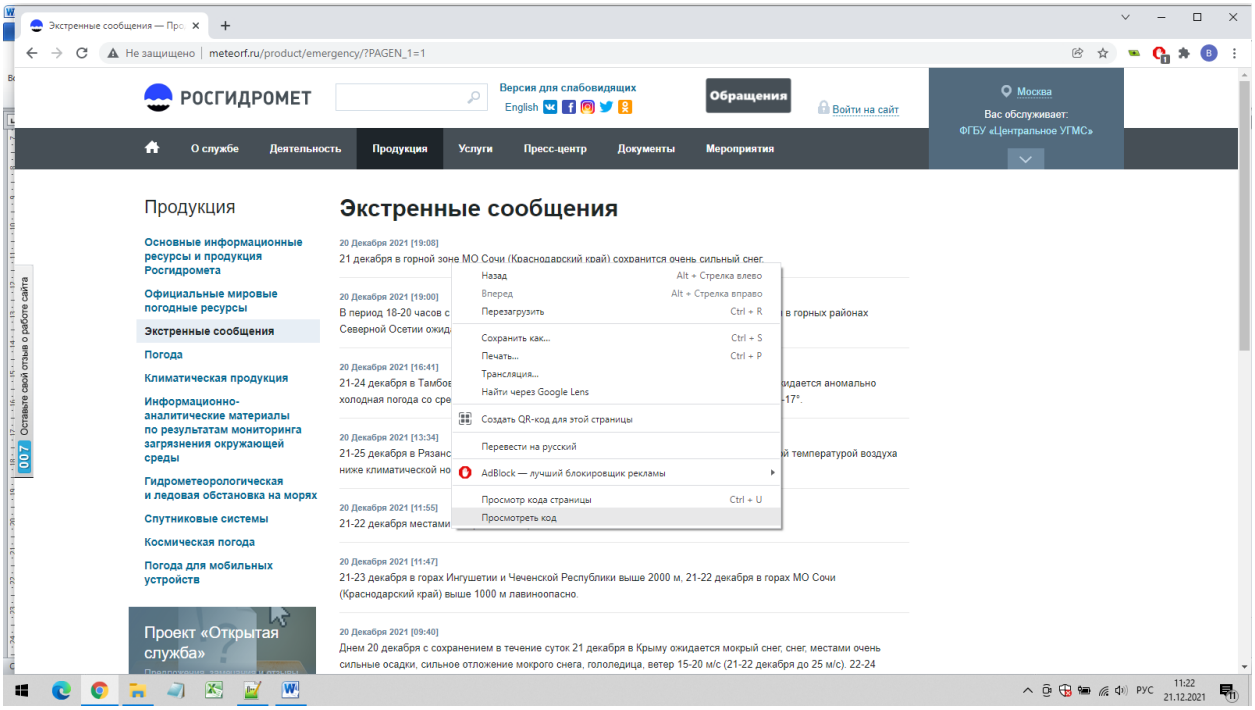
Загрузим страницу с помощью read_html(). С помощью paste() соединим название страницы и нужный номер.

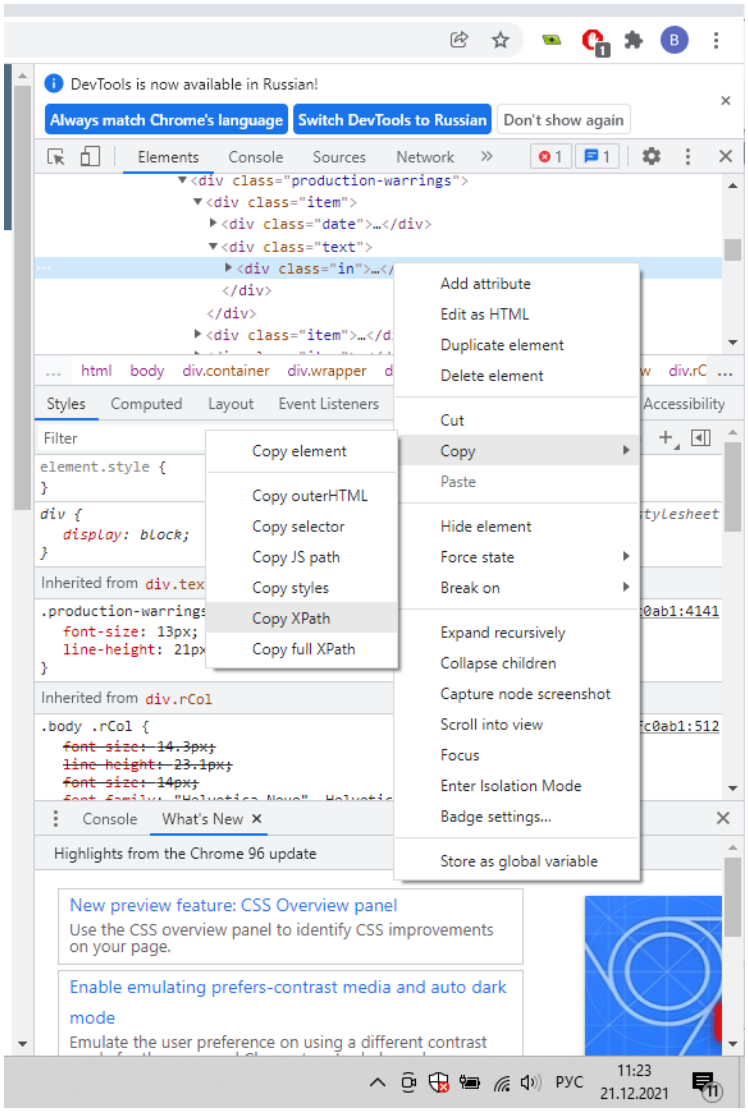
```
i =1
```

```
web_link=paste("http://www.meteorf.ru/product/emergency/?PAGEN_1=",i,sep="")
```

```
meteorf=read_html(web_link)
```

Для поиска нужного элемента из HTLM-документа воспользуемся функцией html_node(). Найдем путь к элементу через XPath при помощи браузера.


Экстренные сообщения — Прс
Не защищено | meteorf.ru/product/emergency/?PAGEN_1=1
РОСГИДРОМЕТ
Версия для слабовидящих
English
Обращения
Войти на сайт
Москва
Вас обслуживает:
ФГБУ «Центральное УГМС»
О службе
Деятельность
Продукция
Услуги
Пресс-центр
Документы
Мероприятия
Продукция
Основные информационные ресурсы и продукция Росгидромета
Официальные мировые погодные ресурсы
Экстренные сообщения
Погода
Климатическая продукция
Информационно-аналитические материалы по результатам мониторинга загрязнения окружающей среды
Гидрометеорологическая и ледовая обстановка на морях
Спутниковые системы
Космическая погода
Погода для мобильных устройств
Проект «Открытая служба»
Экстренные сообщения
20 Декабря 2021 [19:08]
21 декабря в горной зоне МО Сочи (Краснодарский край) сохранится очень сильный снег.
20 Декабря 2021 [19:00]
20 Декабря 2021 [16:41]
20 Декабря 2021 [13:34]
20 Декабря 2021 [11:55]
20 Декабря 2021 [11:47]
21-23 декабря в горах Ингушетии и Чеченской Республики выше 2000 м, 21-22 декабря в горах МО Сочи (Краснодарский край) выше 1000 м лавиноопасно.
20 Декабря 2021 [09:40]
Днем 20 декабря с сохранением в течение суток 21 декабря в Крыму ожидается мокрый снег, снег, местами очень сильные осадки, сильное отложение мокрого снега, гололедица, ветер 15-20 м/с (21-22 декабря до 25 м/с). 22-24
Назад Alt + Стрелка влево
Вперед Alt + Стрелка вправо
Перезагрузить Ctrl + R
Сохранить как... Ctrl + S
Печать... Ctrl + P
Трансляция...
Найти через Google Lens
Создать QR-код для этой страницы
Перевести на русский
AdBlock — лучший блокировщик рекламы
Просмотр кода страницы Ctrl + U
Просмотреть код
Оставьте свой отзыв о работе сайта
РУС
11:22
21.12.2021


DevTools is now available in Russian!
Always match Chrome's language
Switch DevTools to Russian
Don't show again
Elements
Console
Sources
Network

…

…</

…</d
html body div.container div.wrapper d
div.rC
Styles
Computed
Layout
Event Listeners
Accessibility
Filter
element.style {
}
div {
display: block;
}
Inherited from div.tex
.production-warrings
font-size: 13px;
line-height: 21px
}
Inherited from div.rCol
.body .rCol {
font-size: 14.3px;
line-height: 23.1px;
font-size: 14px;
Add attribute
Edit as HTML
Duplicate element
Delete element
Cut
Copy
Paste
Hide element
Force state
Break on
Expand recursively
Collapse children
Capture node screenshot
Scroll into view
Focus
Enter Isolation Mode
Badge settings...
Store as global variable
Copy element
Copy outerHTML
Copy selector
Copy JS path
Copy styles
Copy XPath
Copy full XPath
Console
What's New
Highlights from the Chrome 96 update
New preview feature: CSS Overview panel
Use the CSS overview panel to identify CSS improvements on your page.
Enable emulating prefers-contrast media and auto dark mode
Emulate the user preference on using a different contrast
РУС
11:23
21.12.2021

Получаем такой путь для первого сообщения на странице:

/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[1]/div[2]/div.

Дата первого сообщения:

/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[1]/div[1]/div.

Для второго сообщения дата имеет XPath:

/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[2]/div[1]/div,

а само сообщение:

/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[2]/div[2]/div.

Следовательно, для последнего div[k]: при k=1 – дата сообщения, при k=2 – само сообщение. Для предпоследнего div[j]: j – номер сообщения на странице, от 1 до 10.

Считаем сообщения за 2021 и 2020 годы. Для этого в дате сообщения нужно выделить год.

Получим дату первого сообщения с помощью функций html_node() по XPath и html_text().

```
j=1

k=1

xpath_jk=paste("/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[",j,"]/div[",k,"]/div",sep="")

meteorf_node = html_node (meteorf, xpath=xpath_jk)

date_text=html_text(meteorf_node)

date_text

[1] "\n                    20 Декабря 2021 [19:08]\n                "
```

Разобьем фразу на слова с помощью str_split () из пакета stringr. В качества шаблона, по которому произойдет разбиение, укажем регулярное выражение " +" – хотя бы один пробел.

```
str_split (date_text, " +")
```

В результате получим:

```
[[1]]

[1] "\n"        "20"        "Декабря"   "2021"      "[19:08]\n" ""
```

Нам нужно четвертый элемент полученного списка, то есть:

```
str_split (date_text, " +")[[1]][4]
```

возвращает год.

Запустим считывание все данных со страниц в цикле while, пока год превышает 2019.

В вектор text_all будем записывать текст сообщений, в вектор god_all – годы.

```
text_all = NULL

god_all = NULL

god = 2021

i=1

while (god>2019){

web_link=paste("http://www.meteorf.ru/product/emergency/?PAGEN_1=",i,sep="")

meteorf=read_html(web_link)

for (j in 1:10){

for (k in 1:2){

xpath_jk=paste("/html/body/div[1]/div[1]/div[1]/div/div[3]/div[2]/div/div[1]/div[",j,"]/div[",k,"]/
div",sep="")

meteorf_node=html_node(meteorf,xpath=xpath_jk)

text = html_text(meteorf_node)

if (k == 1) {god = str_split (text, " +")[[1]][4]; god_all = c(god_all, god)}

if (k == 2) text_all = c(text_all, text)

}}

i =i +1

}
```

Проверим, за какие годы считались данные:

```
table(god_all)

god_all

2019 2020 2021

  4   780  1126
```

На последней итерации получили 4 сообщения за 2019 год, далее они игнорируюся.

Воспользуемся пакетом tm, чтобы проанализировать частотное распределение слов в сообщениях за 2019 и 2020 годы.

text_mess = text_all

Представим даные в виде корпуса текста.

docs = Corpus(VectorSource(text_mess))

Преобразуем прописные буквы в строчные.

docs = tm_map(docs, content_transformer(tolower))

Удалим числа.

docs = tm_map(docs, removeNumbers)

Создадим матрицу «документ-термин» (DTM). Она содержит частоту каждого слова (строка) в сообщении (столбец).

dtm = TermDocumentMatrix(docs)

m = as.matrix(dtm)

Отсортируем слова по убыванию частоты встречаемости.

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

v [1:20]

cbind(v [1:20])

[,1]

до 1260

на 950

ожидается 926

местами 874

области 674

суток 661

сильный 623

м/с. 605

ночью 596

ожидаются 536

ветер 464

сентября 454

конца 435

дождь, 408

июля 365

очень 364

августа 364

горах 357

заморозки 353

сильные 342

Сравним, насколько отличается частота слова “заморозки” в 2021 и 2020 году. Для этого в матрице m отберем строку со словом “заморозки”. И применим функцию tapply(), чтобы посчитать сумму частот для каждого года.

```
tapply ( m [ "заморозки",], god_all, sum)

2019 2020 2021
   1  130  222
```

Тем самым абсолютная частота в 2021 году выше, чем в 2020, однако, число сообщений в 2021 году также больше. Рассчитаем долю:

```
tapply ( m [ "заморозки",], god_all, sum) / table(god_all)

2019       2020       2021
0.2500000 0.1666667 0.1971581
```

Оказывается, что 2021 году число случаев, когда упоминалось слово “заморозки”, деленное на число сообщений выше, чем в 2020 г. Но здесь учитывается частота слова “заморозки”, хотя оно может встречаться в одном сообщении несколько раз. Чтобы посчитать долю сообщений со слово заморозки, поменяем в tapply() функцию sum() на function (x) sum(x!=0), посчитывающую число раз, когда частота была не нулевая.

```
tapply ( m [ "заморозки",], god_all, function (x) sum(x!=0))/ table(god_all)

2019       2020       2021
0.2500000 0.1487179 0.1927176
```

Как видим, доля сообщений об экстремальных погодных явлениях с упоминанием заморозков в 2021 году выше, чем в 2019. Хотя в выборку 2021 года не вышли последние 10 дней декабря.

Проверим, значимы ли эти различия с помощью критерия Стьюдента.

Для этого создадим бинарные векторы (0 – нет слова “заморозки” в сообщении), 1 – иначе), отдельно за 2021 и 2020 годы.

```
v_2021 = ifelse (m [ "заморозки",] [god_all == "2021"] ==0, 0, 1)

v_2020 = ifelse (m [ "заморозки",] [god_all == "2020"] ==0, 0, 1)

t.test(v_2021, v_2020)

        Welch Two Sample t-test

data:  v_2021 and v_2020

t = 2.5369, df = 1777.6, p-value = 0.01127

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 0.009983196 0.078016075

sample estimates:

mean of x mean of y

0.1927176 0.1487179
```

Оказалось, что наблюдается значимая на 95%-м доверительном уровне разница между упоминаниями слова “заморозки” в экстренных сообщениях в 2021 и 2020 годах.