# Парсер для новостных сайтов.

Скрипт, который парсит новостные сайты. Скрипт является универсальным, не меняя внутренности кода, он может парсить **ЛЮБЫЕ\*** сайты, которые будут лежать в таблице `resource`.

Список сайтов которые можно запарсить (Можете брать любые сайты, на ваще же усмотрение ):

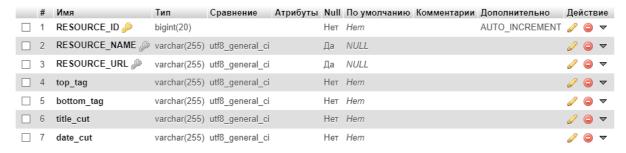
- 1. Nur.kz
- 2. Scientificrussia

## Детали парсинга

Каждый сайт - отдельный ресурс в моей таблице. Заходит по ссылке на ресурс, чтобы вытащить ссылки на все новости, и спарсить новости. От каждой новости забираю 4 важных атрибута\*: ссылку, заголовок, контент и дату.

# Таблица resource

Таблица **resource** хранит ресурсы для парсинга. В этой таблице хранятся ресурсы(новостные сайты) по которым парсер должен собрать новости. Имеет следующую структуру:



#### Поля:

**RESOURCE\_ID** - Это поле автоматически генерирует уникальный номер для ресурса в таблице.

**RESOURCE\_NAME** - Это поле содержит краткое название ресурса.

RESOURCE URL - Это поле содержит ссылку на ресурс где парсер забирает новости.

top\_tag - Это поле содержит структуру для взятие ссылок на новости.

**bottom\_tag** - Это поле содержит структуру для взятия текстового контента новости.

title\_cut - Это поле содержит структуру для взятия заголовка новости.

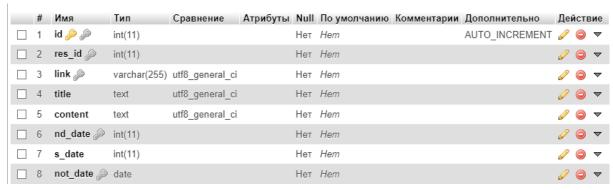
date\_cut - Это поле содержит структуру для взятия даты и времени новости.

### Пример записи в таблице resource:



# Таблица items

Таблица **items** хранит новости которые успешно собрал парсер. Имеет следующую структуру:



### Поля:

id - Это поле автоматически генерирует уникальный номер для новости в таблице.

res id - это поле resource id из таблицы resource.

link - это поле содержит ссылку на новость.

title - это поле содержит заголовок новости.

content - это поле содержит текстовый контент новости.

nd\_date - это поле содержит дату и время новости в формате Unix time.

**s\_date** - это поле содержит дату и время попадания новости в саму таблицу **items** в формате **Unix time**.

not\_date - это поле содержит дату новости в формате Год-Месяц-День.

#### Пример записи в таблице **items**:

