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:

#	Имя	Тип	Сравнение	Атрибуты	Null	По умолчанию	Комментарии	Дополнительно	Действие
1	RESOURCE_ID	bigint(20)			Нет	*Нет*		AUTO_INCREMENT	
2	RESOURCE_NAME	varchar(255)	utf8_general_ci		Да	*NULL*			
3	RESOURCE_URL	varchar(255)	utf8_general_ci		Да	*NULL*			
4	top_tag	varchar(255)	utf8_general_ci		Нет	*Нет*			
5	bottom_tag	varchar(255)	utf8_general_ci		Нет	*Нет*			
6	title_cut	varchar(255)	utf8_general_ci		Нет	*Нет*			
7	date_cut	varchar(255)	utf8_general_ci		Нет	*Нет*			

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**:

RESOURCE_ID	RESOURCE_NAME	RESOURCE_URL	top_tag	bottom_tag	title_cut	date_cut
1	Информационный портал ZAKON.KZ	https://www.zakon.kz	структура для взятия ссылок	структура для взятия текстового контента новости	структура для взятия заголовка новости	структура для взятия даты и времени новости

Таблица items

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

#	Имя	Тип	Сравнение	Атрибуты	Null	По умолчанию	Комментарии	Дополнительно	Действие
1	id	int(11)			Нет	*Нет*		AUTO_INCREMENT	
2	res_id	int(11)			Нет	*Нет*			
3	link	varchar(255)	utf8_general_ci		Нет	*Нет*			
4	title	text	utf8_general_ci		Нет	*Нет*			
5	content	text	utf8_general_ci		Нет	*Нет*			
6	nd_date	int(11)			Нет	*Нет*			
7	s_date	int(11)			Нет	*Нет*			
8	not_date	date			Нет	*Нет*			

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**:

id	res_id	link	title	content	nd_date	s_date	not_date
1	777	Ссылка на новость	Заголовок новости	Текстовый контент новости	1629793430	1629793431	2021-08-24