**Министерство науки и высшего образования Российской Федерации** **Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет** **имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

**Факультет «Информатика и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Проект BMSTU-NEWS

по дисциплине «Программирование сетевых приложений»

Выполнил:

студент группы ИУ5-43Б

Некрасов С. А.

Проверила:

Аксёнова М. В.

2022 г.

**Описание задания:**

Создать сервис парсинга новостей с официального сайта МГТУ.

Примечание: весь код проекта выгружен на репозиторий по ссылке

<https://github.com/Nekrasov-Sergey/bmstu-news>

**Layout проекта:**

Проект создан по общепринятому макету для проектов приложений на языке Golang.

* **cmd/**

В папке cmd/ содержится основное приложение проекта. В папке cmd/bmstu-news содержится main.go файл, который использует и вызывает код из папок internal/ и pkg/. В папке cmd/first-parse храниться код для загрузки в базу данных всех доступных новостей с сайта МГТУ. В папке cmd/migrate содержится код для создания миграций для базы данных.

* **config/**

В папке config/ содержатся все основные конфиги проекта.

* **internal/**

В папке internal/ содержится внутренний код проекта, таким образом мы показываем, что данный код не для импортирования в другие приложения или библиотеки. Сам код приложения находится в папке internal/app. В папке internal/pkg содержится код, используемый приложением.

* **internal/app**

В папке app/config хранится файл config.go, где создается структура конфигурации, которая содержит все конфигурационные данные о сервисе. В файле context.go содержится код для возвращения конфига из контекста, дополнения контекста конфигом.

В папке app/dsn хранится файл dsn.go, в котором содержится код для сбора DSN строки из переменных окружения для подключения к базе данных.

В папке app/model хранится файл model.go, в котором содержатся структуры для записи новостей, структура NewsItem содержит данные короткого представления новости, структура FullNewsItem содержит данные полного представления новости, структура News получается из объединений структур NewsItem и FullNewsItem, она предназначена для записи новости в базу данных.

В папке app/repository хранится файл repo.go, в котором содержится код для подключения к базе данных и записи в нее новости.

В папке app/service хранится файл service.go, в котором содержится структура service и функции данной структуры. Здесь содержится код для парсинга двух представлений новости, объединения этой информации в общую структуру.

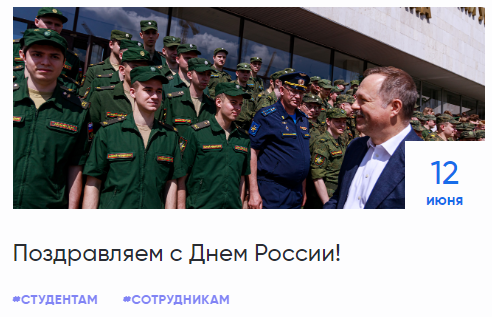
* **internal/pkg**

В папке pkg/app хранится файл app.go, где содержится структура приложения, основные функции создания приложения, запуска и записи всех новостей.

В папке pkg/clients/news хранится клиент news. В файле client.go содержатся основные функции для «сырого» запроса двух представлений новости. Также здесь содержатся файлы, содержащие структуры для записи первичного представления новостей с сайта.

**Описание работы сервиса:**

На сайте мгту новости представлены в двух видах – это короткий вид и полный, когда мы открываем конкретную новость.



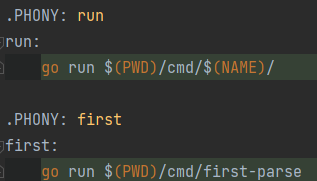
*Короткий вид новости*



*Часть полного вида новости*

Часть важных данных о новости хранится в коротком представлении, другая часть – в полном, поэтому мы будем делать два запроса для каждого представления новости, объединять эти данные в одну общую структуру и записывать в базу данных.

Для удобства запуска различных функций сервиса будем использовать makefile, где будем прописывать команды, например команды запуска и записи всех новостей:



Парсинг двух представлений новостей содержится в функциях клиента GetNews и GetFullNews. Они используются в функциях сервиса ParseShortNews и ParseFullNews, где данные с сайта мы записываем в более удобные и правильные для нас структуры.

Приложение при запуске с определенной периодичностью запрашивает с сайта определенное в конфиге количество новостей, проверяет эти новости на соответствие с новостями, уже записанными в базу данных, и производит запись.

Для записи в базу данных используется Docker, в docker-compose файле мы прописываем параметры создаваемого контейнера и базы данных.

Также для первого запуска приложения был создан функционал по записи абсолютно всех новостей в базу данных. Такая запись реализована в функции приложения FirstParse с помощью горутин, которые позволяют параллельно производить запросы и запись в базу данных, тем самым позволяя эффективно и быстро заполнить базу данных почти пятью тысячами новостей.

**База данных:**

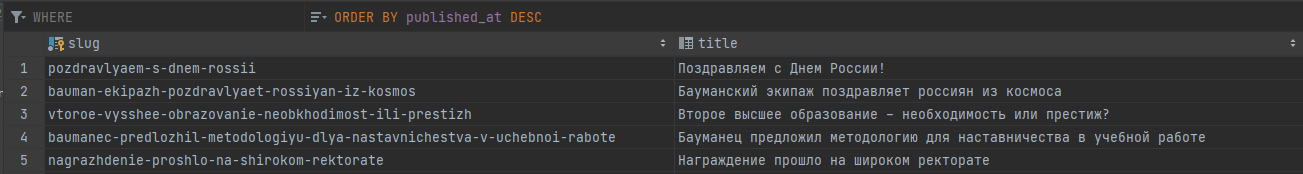
Каждая новость в базе данных представлена в виде одной строки. Для создания и записи базы данных был использован PostgreSQL.

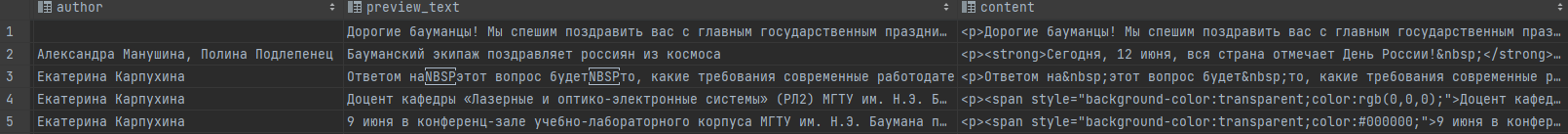
Рассмотрим, как выглядит новость в базе данных, в таблице news содержатся следующие атрибуты:



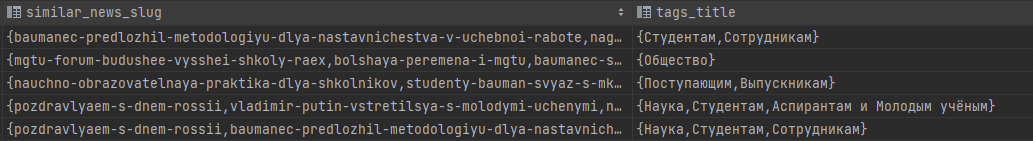
Таким способом, в поле slug записывается идентификатор новости, в поле title ее название, в поле author автор новости, в поле preview\_text превью запись новости, в поле content вся текстовая информация внутри новости, в поле reading\_time время прочтения новости, в поле published\_at дата публикации, в поле image картинка-превью новости, в поле photo\_report опционально, в зависимости от содержания в новости, записывается массив картинок, которые листаются в конце новости, в поле similar\_news\_slug хранится массив с идентификаторами похожих новостей, в поле tags\_title хранится массив с тэгами новости, по которым все новости можно будет отфильтровать (пример тэга – Обучение, Наука).

Вот так выглядят новости в базе данных:

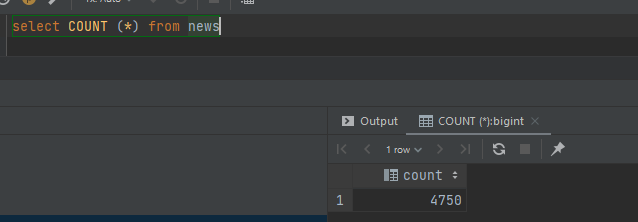


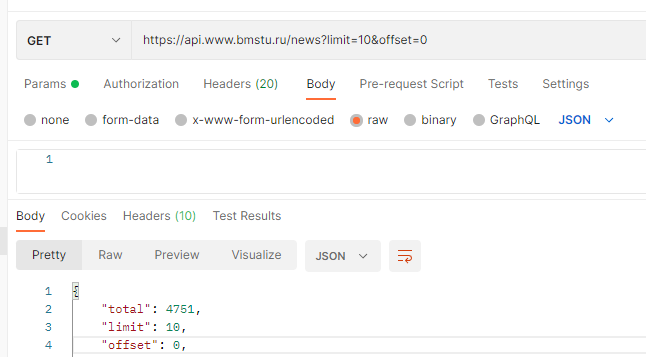




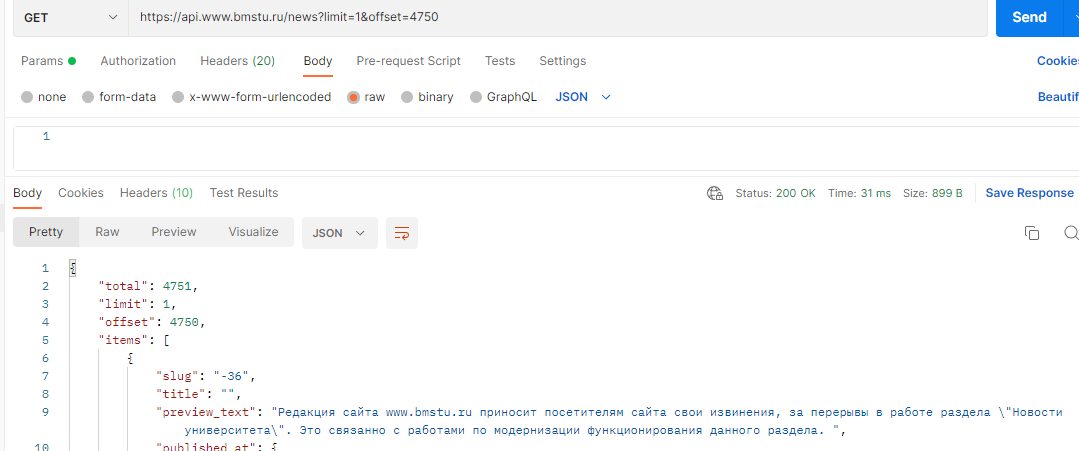


Произведем запуск парсинга всех новостей и проверим, сколько новостей будет записано в базу данных:



Как мы видим, было записано 4750 новостей, проверим это число через прямой запрос к сайту МГТУ:  


Расхождение в одну новость вызвано тем, что 4751 новость недоступна, попробуем ее запросить:



Таким образом нами был создан сервис, позволяющий запрашивать новости с сайта МГТУ, преобразовывать их и записывать в базу данных.