Crypto Project. Сборщик новостей. Техническое задание

1. Общая информация

   Есть два новостных сайта cryptocoinsnews.com и cointelegraph.com. Наша задача выгрузить все имеющиеся новости с этих сайтов, затем с некоторой периодичностью проверять появление новой информации и тоже ее выгружать.

   Загруженную новость мы должны смапить с нашей моделью данных, проверить на дубликаты и сохранить в нашем хранилище.

   Проект должен быть написан на C# 6.0+, .Net Framework 4.6+

2. Структура новостных сайтов

   На указанных выше сайтах информации больше чем нам нужно, поэтому в этом разделе будет указано что именно и откуда мы будем качать.

   Cryptocoinsnews

   Новость на этом ресурсе помечается одним или несколькими тегами. После этого новость будет доступна в той категории, тег которой был задан. Если будет задано несколько тегов - новость будет отображаться в нескольких категориях.

   Есть тег News он представляет глобальную категорию для всех новостей. Проблема в том что не все новости помечаются этим тегом. Поэтому выгружать новости необходимо из каждой категории в отдельности исключая при этом повторы, которых будет не мало.

   Список категорий и ссылок:

News	https://www.cryptocoinsnews.com/news/
Bitcoin Politics	https://www.cryptocoinsnews.com/bitcoin-politics/
Bitcoin Mining	https://www.cryptocoinsnews.com/bitcoin-mining/
Bitcoin Analysis	https://www.cryptocoinsnews.com/bitcoin-analysis/

Bitcoin Price News	https://www.cryptocoinsnews.com/bitcoin-price-news/
Altcoin News	https://www.cryptocoinsnews.com/altcoin-news/
ICO	https://www.cryptocoinsnews.com/ico/
Blockchain News	https://www.cryptocoinsnews.com/blockchain-news/
Bitcoin Opinion	https://www.cryptocoinsnews.com/bitcoin-opinion/
Bitcoin Technology	https://www.cryptocoinsnews.com/bitcoin-technology/
FinTech News	https://www.cryptocoinsnews.com/fintech/
Bitcoin Regulation	https://www.cryptocoinsnews.com/bitcoin-regulation/
Bitcoin Progress	https://www.cryptocoinsnews.com/bitcoin-progress/
Bitcoin Exchange	https://www.cryptocoinsnews.com/bitcoin-exchange-news/
Hacked	https://www.cryptocoinsnews.com/hacked-2/
Ethereum News	https://www.cryptocoinsnews.com/ethereum-news/
Bitcoin Education	https://www.cryptocoinsnews.com/bitcoin-education/
Banking	https://www.cryptocoinsnews.com/banking/
Sponsored Stories	https://www.cryptocoinsnews.com/sponsored-stories/
Bitcoin Business	https://www.cryptocoinsnews.com/bitcoin-business/
Bitcoin Announcements	https://www.cryptocoinsnews.com/bitcoin-announcements/
Altcoin Mining	https://www.cryptocoinsnews.com/cryptocur

	rency-mining/
Guest Contributor	https://www.cryptocoinsnews.com/guest-contributor/

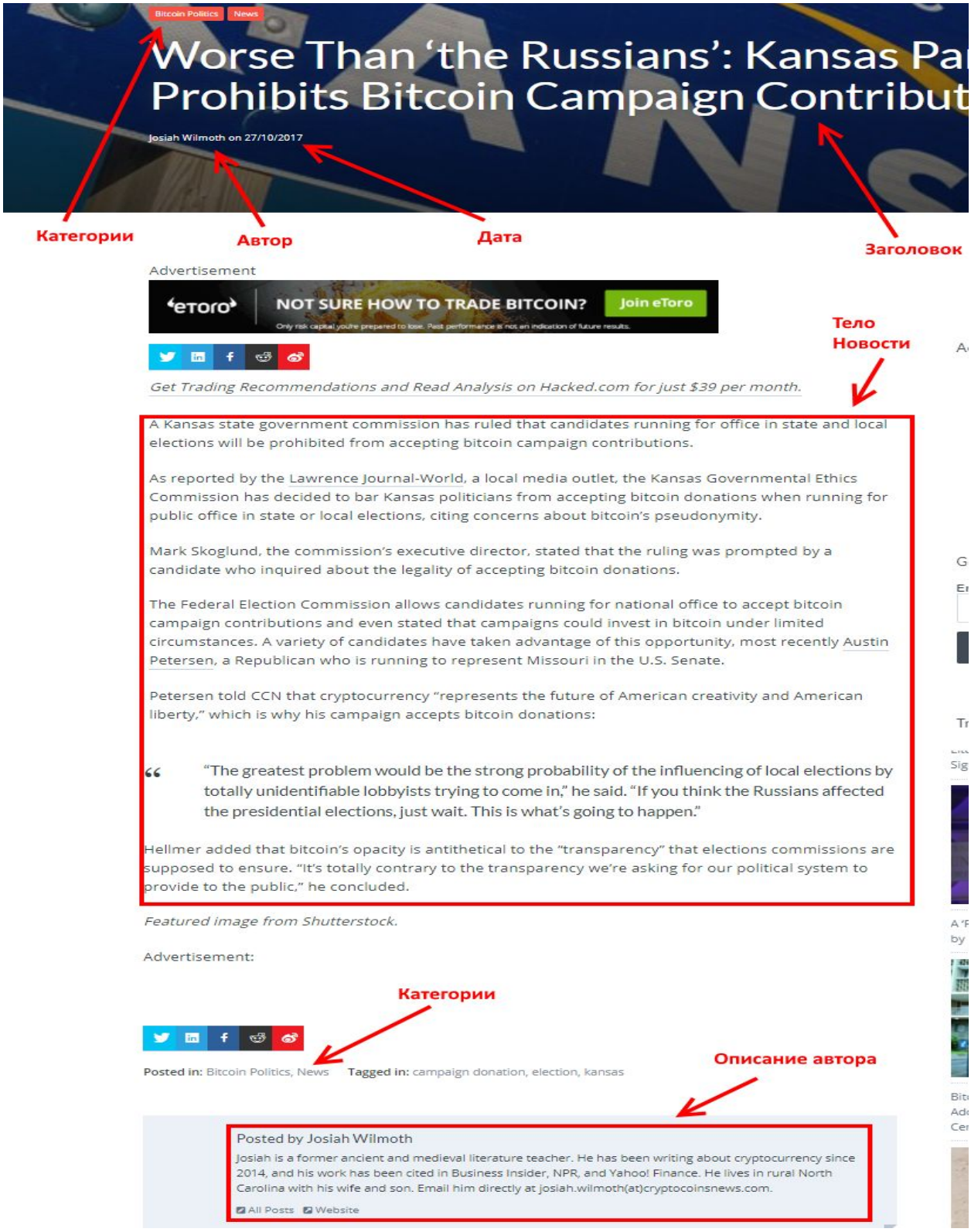
Список категорий не полный. Вам необходимо самостоятельно его дополнить.

Должна быть возможность задать в конфигурации список категорий, которые необходимо исключить (не загружать). Пример: есть категория Sponsored Stories, которая явно не нужна.

Нужно обратить внимание, что новости в категории разбиты на несколько страниц. К примеру: в категории News 606 страниц. Так же и в других категориях. Необходимо выгружать новости со всех страниц в каждой категории.
Пример ссылки на вторую страницу категории News:
https://www.cryptocoinsnews.com/news/page/2/

Новости в категории

Страница новости



Какую информацию собираем с cryptocoinsnews.com?

- Ссылка на страницу новости
- Заголовок.
- Список всех категорий
- Автор
- Дата. Внимание, здесь дата представлена только днем (27/10/2017) без указания времени. Выгружая историю новостей мы берем только дату, время ставим 00:00. При загрузке обновлений ставить то время, когда новость была обнаружена программой (только при условии того же дня). К примеру, мы выкачиваем новости сегодня (28 октября), видим что поступила новости за 28 октября, ставим для новости текущее время (таймзона EST).
- Тело новости

Cointelegraph

Как и у cryptocoinsnews, здесь новости помечаются тегами и распределяются по категориям. Необходимо выгружать все новости со всех категорий.

Список категорий и ссылок:

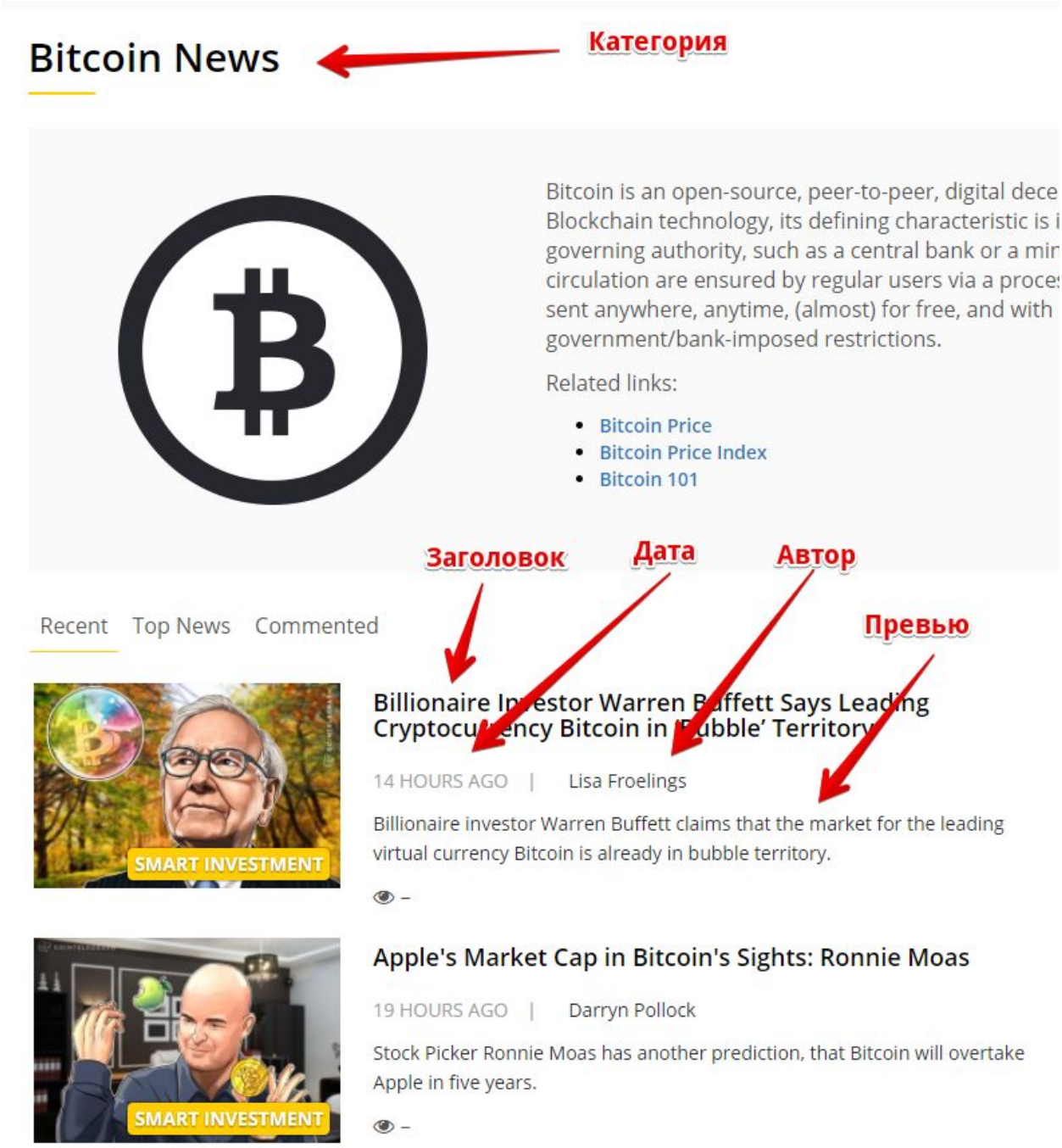
Blockchain News	https://cointelegraph.com/tags/blockchain
Banks News	https://cointelegraph.com/tags/banks
Bitcoin News	https://cointelegraph.com/tags/bitcoin
Bitcoin Price News	https://cointelegraph.com/tags/bitcoin-price
Tradings	https://cointelegraph.com/tags/tradings
Bitcoin Exchanges News	https://cointelegraph.com/tags/bitcoin-exchanges
Bitcoin Wallet News	https://cointelegraph.com/tags/bitcoin-wallet
Bitcoin News	https://cointelegraph.com/tags/bitcoin
Bitcoin Mining News	https://cointelegraph.com/tags/bitcoin-mining

Список категорий не полный. Вам необходимо самостоятельно его дополнить.

Должна быть возможность задать в конфигурации список категорий, которые необходимо исключить (не загружать).

Нужно обратить внимание, что новости в категории отображаются не все, их нужно подкачивать нажимая на кнопку “Load more articles”.

Новости в категории



Страница новости

Какую информацию собираем с cointelegraph.com?

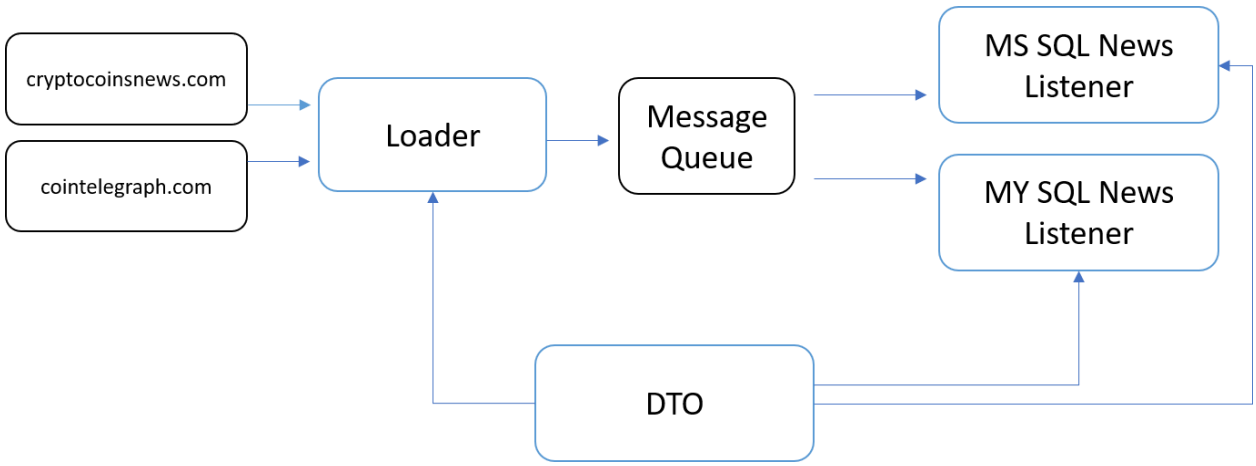
- Ссылка на страницу новости
- Заголовок.
- Список всех категорий
- Автор
- Дата. Внимание, здесь дата бывает указана в нескольких форматах: “15 HOURS AGO” или “OCT 27, 2017”.

  В первом случае необходимо от текущего времени EST отнять указанное кол-во часов (минут и т д).

  Во втором случае, выгружая историю новостей мы берем только дату, время ставим 00:00. При загрузке обновлений ставить то время, когда новость была обнаружена программой (только при условии того же дня). К примеру, мы выкачиваем новости сегодня (28 октября), видим что поступила новости за 28 октября, ставим для новости текущее время (таймзона EST).
- Тело новости

3. Структура проекта

Мы видим реализацию проекта с использованием микросервисной архитектуры. Передача данных между компонентами системы должно быть реализовано с помощью очереди сообщений (мы используем RabbitMq).



Задача реализовать компоненты обозначенные на схеме выше синим цветом.

Loader

Это windows service hosted as console app (испольщуется Topshelf http://topshelf-project.com/), которая при первом запуске выгружает с указанных сайтов историю новостей (должна быть настройка в конфигурации: грузить ли историю).

Загрузив историю, программа должна с заданной периодичностью (настройка в конфигурации) выгружать новые данные (новости).

История новостей загружаться пачками заданного объема. Объем пачки должен задаваться в конфигурации. К примеру, 100 шт. Это значит, что скачивая историю мы будем вытаскивать 100 новостей, отдавать их в очередь, а затем закачивать следующие 100.

Под выгрузкой новых данных (новостей) понимается, что программа будет выгружать только последние X новостей (X задается в конфигурации, к примеру 10) из каждой категории и передавать их в очередь сообщений.

Проблема дубликатов
Принимая во внимание сказанное выше у нас будет проблема с дубликатами, как внутри одного новостного сайта, так и дубликатами обоих сайтах.

Было бы логичным ориентироваться на время новости, но это невозможно поскольку используемые новостные сайты не предоставляют точное время выхода новости, только день.

Мы хотим решить проблему дубликатов использую ссылки на страницы с новостями. В базе не должно быть двух записей ссылающихся на одну и туже страницу с новостью. Мы реализуем эту проверку самостоятельно на уровне базы данных.

Это решение, к сожалению, не поможет нам в том случае, когда одна и та же новость размещена сразу на двух сайтах.

Мы готовы рассмотреть ваши предложения по решению проблемы дубликатов.

MS SQL News Listener и My SQL New Listener
Это windows service hosted as console app (испольщуется Topshelf http://topshelf-project.com/). Нам нужно реализовать две службы, которые будут принимать новости из очереди и добавлять их в соответствующую базу данных.

Добавлять информацию в базу нужно не напрямую в таблицу, а с использованием хранимых процедур. Сигнатуры будут предоставлены позднее.

DTO и очередь сообщений
В качестве очереди сообщений мы используем RabbitMq. Для упрощения

взаимодействия с RabbitMq мы используем библиотеку EasyNetQ (http://easynetq.com/). С ее помощью мы можем передавать типы Net от поставщика к потребителю.

Чтобы это заработало нам необходимо поместить DTO типы в отдельную сборку и использовать ее во всех проектах Loader и MS SQL News Listener, My SQL New Listener.

В нашем случае в сборке должен быть определен только один тип News (подробнее дальше).

4. Наша модель данных

```
enum ProviderType
{
  cryptocoinsnews,
  cointelegraph
}

interface News
{
  Guid ID { get; set; }
  DateTime CreatedTime { get; set; }
  ProviderType Provider { get; set; }
  string Link { get; set; }
  DateTime Date { get; set; }
  IEnumerable<string> Categories { get; set; }
  string Author { get; set; }
  string Header { get; set; }
  string Body { get; set; }
}
```

5. Параметры конфигурации

Просим вас вынести в конфигурацию любые настройки, которые можно было бы изменить и которые могут повлиять на работу программы.

Как минимум для проекта Loader:
- грузить ли историю (по умолчанию false);
- название категорий, через запятую, которые не нужно загружать (по умолчанию пусто);
- настройки Topshelf;
- частота загрузки новых данных (по умолчанию 1 мин)

- размер пачки исторических данных (по умолчанию 100)
- размер пачки новых данных (по умолчанию 10)
- строка подключения EasyNetQ

Как минимум для проектов MS SQL News Listener, My SQL New Listener:
- настройки Topshelf;
- строка подключения EasyNetQ;
- строка подключения к базе данных

6. Тестовое окружение

   После начала работы над проектом мы подготовим тестовое окружение (базы данных, очередь сообщений) и дадим вам доступ для тестирования.

7. Языки и технологии

   C# 6.0+
   .Net Framework 4.6+
   Visual Studio 2015

   Пространство имен по умолчанию во всех проектах: cryptofund

   Для создания windows service использовать topshelf. Настройки topshelf выводятся в конфигурацию.

   Для логирования использовать .NLog. Конфигурацию логгера поместить в NLog.config

   Очередь сообщений: RabbitMq
   Библиотека для работы с очередью сообщений: http://easynetq.com/

   Базы данных: MS SQL и MY SQL

8. Требования к исходному коду

   Программа должна быть составлена с использование принципов ООП и современных практик в разработке программного обеспечения.

   Код должен быть аккуратно оформлен, хорошо структурирован, иметь понятное наименование и предназначение.

   Настоятельно просим вас принять во внимание, что с составленной вами программой в дальнейшем, достаточно продолжительное время, будут работать другие люди.

9. Что должно быть в результате?

   После завершения работы над проектом вы должны предоставить нам исходные коды проектов (Loader, MS SQL News Listener, My SQL New Listener и DTO library) и сопутствующие материалы, если таковые будут.

   Нам потребуется некоторое время на развертывание и тестирование (1 - 2 дня). Вы обязуетесь исправить найденные нами ошибки, недочеты и реализовать какие-то небольшие пожелания, если они будут.

   Все исключительные и иные права на данную разработку и все материалы по проекту принадлежат заказчику в полном объеме и без ограничений.