

Генератор парсеров

Кормышов Михаил

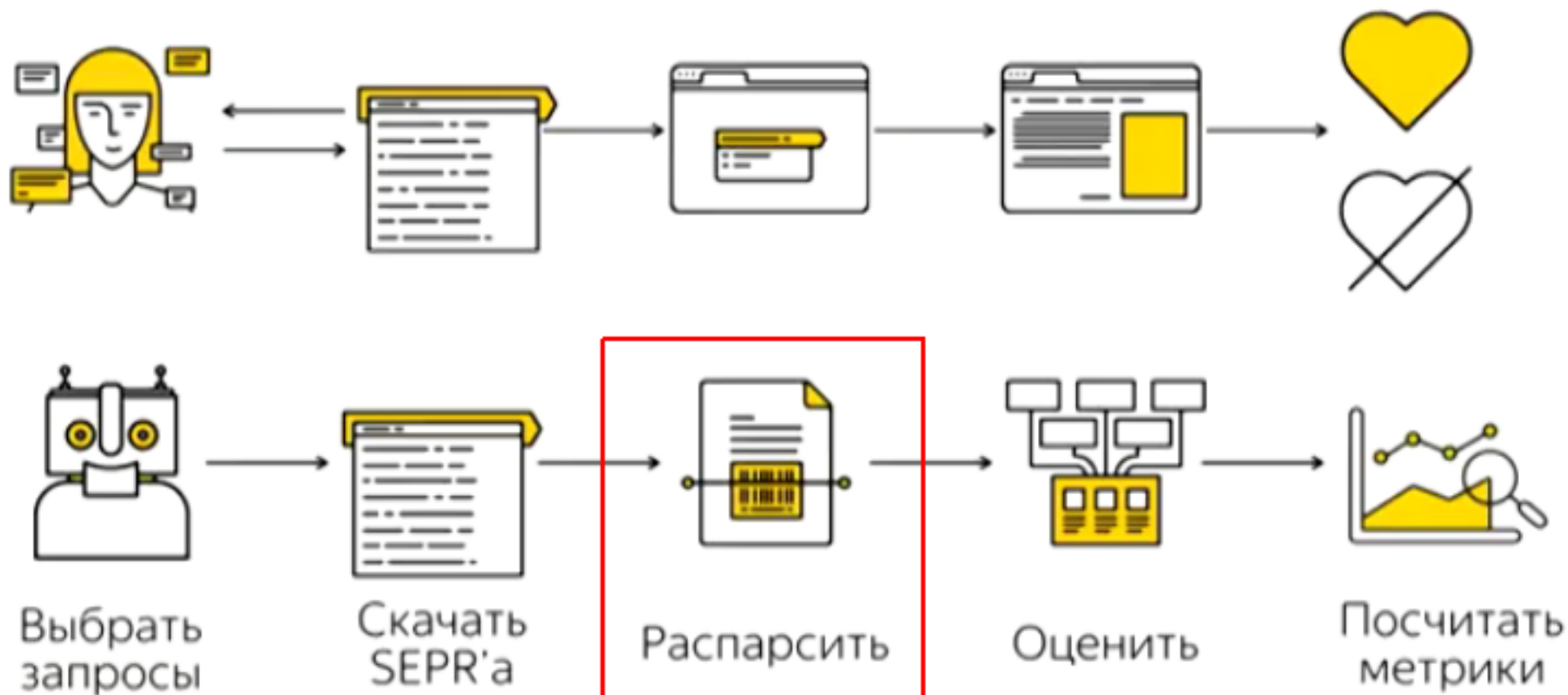
Руководитель: Фёдор Амосов

CSC, осень 2017

Offline оценка качества поиска



Offline оценка качества поиска



Что парсим

Яндекс

карта метро



Найти

ПОИСК

КАРТИНКИ

ВИДЕО

КАРТЫ

МАРКЕТ

НОВОСТИ

ПЕРЕВОДЧИК

Карта метро Санкт-Петербурга - интерактивная схема...

[metrobook.ru](#) > [ru_spb](#)

Для того, чтобы вам легче было сориентироваться и рассчитать время в пути, соз наша интерактивная карта/схема метро Санкт-Петербурга.



Карта метро Санкт-Петербурга

[metromap.ru](#) > [spb/](#)

Карта метро Санкт-Петербурга. Metromap.ru — это интерактивная схема метро, которая поможет вам выбрать оптимальный маршрут и рассчитать время в пути.



карта метро — смотрите картинки

[yandex.ru/images](#) > [карта метро](#)



Официальный сайт Петербургского Метрополитена

[metro.spb.ru](#) > [map.html](#)

Перспективная схема метрополитена. История метро в схемах. Схема метрополитена. © 2007—2017 «Петербургский Метрополитен».

"components": [

{

"page_url": "http://metrobook.ru/ru_spb",

"snippet": "Для того, чтобы вам легче было сорир

"title": "Карта метро Санкт-Петербурга - интера

"type": "SEARCH_RESULT",

"view_url": "http://metrobook.ru/ru_spb"

},

{

"page_url": "http://www.metromap.ru/spb/",

"snippet": "Карта метро Санкт-Петербурга. Metro

"title": "Карта метро Санкт-Петербурга",

"type": "SEARCH_RESULT",

"view_url": "http://www.metromap.ru/spb/"

},

{

"media_links": [

"im0-tub-ru.yandex.net/i?id=fa410d095417110

"im0-tub-ru.yandex.net/i?id=d319e5485287280

"im0-tub-ru.yandex.net/i?id=2aecca2a5903dbf

"im0-tub-ru.yandex.net/i?id=21df03e497d9641

"im0-tub-ru.yandex.net/i?id=4df339c62f9cd3a

],

"page_url": "http://yandex.ru/images/search?text=%D0

"title": "карта метро — смотрите картинки",

"type": "WIZARD",

"wizard_type": "WIZARD_IMAGE"

},

{

"page_url": "http://www.metro.spb.ru/map.html",

"snippet": "Перспективная схема метрополитена. I

"title": "Официальный сайт Петербургского Метро

"type": "SEARCH_RESULT",

"view_url": "http://www.metro.spb.ru/map.html"

Как приходится парсить сейчас

```
wizard.type = "WIZARD"  
wizard.wizard_type = "WIZARD_NEWS"  
wizard.alignment = "LEFT"  
wizard.page_url = self.get_from_page(element, ".", "href")  
wizard.title = self.get_from_page(element, ".", "string")  
return wizard
```

```
def extract_markup(self, file_name):  
    with open(file_name, "r") as file:  
        tree = html.document_fromstring(file.read())  
    markup = SearchMarkup()  
    markup.file = file_name.split('/')[-1]  
    markup.type = "HTMLTree"  
    block_list = tree.xpath("//html/body/table/tbody/tr/td/div/div")  
    for block in block_list:  
        adv_list = block.xpath(".")  
        for adv in adv_list:  
            subblock_list = adv.xpath("./ol/li")  
            for subblock in subblock_list:  
                if len(subblock.xpath("./h3")) > 0:  
                    result = self.extract_adv(subblock)  
                    markup.add(result)  
            subblock_list = block.xpath("./div/div/ol/div")  
            for subblock in subblock_list:  
                document_list = subblock.xpath(".")  
                for document in document_list:  
                    if len(document.xpath("./div/div/cite")) > 0:  
                        result = self.extract_search_result(document)  
                        markup.add(result)  
            wizard_image_list = subblock.xpath(".")  
            for wizard_image in wizard_image_list:  
                if len(wizard_image.xpath("./div[1]/a")) > 0:  
                    result = self.extract_wizard_image(wizard_image)  
                    markup.add(result)  
            wizard_news_list = subblock.xpath("./div/h3/a")  
            for wizard_news in wizard_news_list:
```

Хотим генерировать парсеры

Google

маск

Войти

Все Новости Картинки Видео Карты Ещё Настройки Инструменты

Результатов: примерно 4 870 000 (0,58 сек.)

Маск, Илон — Википедия
https://ru.wikipedia.org/wiki/Маск_Илон
Илон Рив Маск (англ. Elon Reeve Musk /iːlɒn ˈmʌsk/; род. 28 июня 1971 года, Претория, ЮАР) — канадско-американский инженер, предприниматель, филантроп, изобретатель и инвестор; миллиардер. Основатель компании PayPal; основатель, совладелец, генеральный директор и главный инженер ...
Херд, Эмбер · Tesla · SpaceX · Neuralink

Главные новости

Истории Бывший рекрутер Tesla рассказал о том, как Илон Маск набирает...
Rusbase
2 дня назад

Илон Маск запустит в космос автомобиль
Gismeteo
2 дня назад

Компания Элона Маска показала карту тоннелей под Лос-Анджелесом
Популярная механика
1 день назад

→ Больше результатов по запросу "маск"

Илон Маск / Elon Musk (Tesla, SpaceX, Neuralink) | ВКонтакте
<https://vk.com/elonmusk>
неофициальная страница* Илона Маска (также Элон Маск, англ. Elon Musk)

Падение Илона Маска началось: СМИ наконец подсчитали его ...
<https://da.ru/analytics/20171107/1508279232.html>
7 нояб. 2017 г. - Текущие события вокруг, возможно, самого известного бизнесмена в мире Илона Маска, события, связанные с его деятельностью, его личными делами, финансовыми результатами...

SEARCH_RESULT
Title
URL
Description

WIZARD_NEWS
URL

Аналоги

ScraperLab



Алгоритм

Google

маск




Войти

Все Новости Картинки Видео Карты Ещё Настройки Инструменты

Результатов: примерно 4 870 000 (0,58 сек.)

Маск, Илон — Википедия
https://ru.wikipedia.org/wiki/Маск,_Илон
Илон Рив Маск (англ. Elon Reeve Musk /iːlɒn ˈmʌsk/; род. 28 июня 1971 года, Претория, ЮАР) — канадско-американский инженер, предприниматель, филантроп, изобретатель и инвестор; миллиардер. Основатель компании PayPal; основатель, совладелец, генеральный директор и главный инженер ...
Хёрд, Эмбер · Tesla · SpaceX · Neuralink

Главные новости

 <p>Истории Бывший рекрутер Tesla рассказал о том, как Илон Маск набирает...</p> <p>Rusbase</p> <p>2 дня назад</p>	 <p>Илон Маск запустит в космос автомобиль</p> <p>Gismeteo</p> <p>2 дня назад</p>	 <p>Компания Элона Маска показала карту тоннелей под Лос-Анджелесом</p> <p>Популярная механика</p> <p>1 день назад</p>
---	--	---

→ Больше результатов по запросу "маск"

Илон Маск / Elon Musk (Tesla, SpaceX, Neuralink) | ВКонтакте
<https://vk.com/elonmusk>
неофициальная страница* Илона Маска (также Элон Маск, англ. Elon Musk)

Падение Илона Маска началось: СМИ наконец подсчитали его ...
<https://ria.ru/analytics/20171107/1508279232.html>
7 нояб. 2017 г. - Текущие события вокруг, возможно, самого известного бизнесмена в мире Илона Маска приближены к пикову срывая на экономических пятаках и финансовыми показателями

Diagram illustrating the algorithm structure:

```
graph LR; SR1[SEARCH_RESULT] -.-> Node1(( )); SR2[SEARCH_RESULT] -.-> Node1; Node1 -- "/html/body/ ..." --> Node2(( )); Node2 -- ".div" --> Node3(( )); Node3 -- ".lg-section-img/ ..." --> WIZARD_IMAGE[WIZARD_IMAGE]; Node3 -- ".div/div" --> WIZARD_NEWS[WIZARD_NEWS];
```


Алгоритм



репетитор



Войти

Все

Картинки

Новости

Видео

Карты

Ещё

Настройки

Инструменты

Результатов: примерно 6 240 000 (0,54 сек.)

Подбор репетиторов на PROFI.RU - Проверенные специалисты

Реклама spb.profi.ru/Репетиторы/в_Петербурге ▼

Опытные **репетиторы** на PROFI.RU. Более 15 000 анкет. Бесплатный и быстрый подбор!

Занятия в группе · 16 689

Курсы: Английский язык, Математика, Русский язык, Обществознание

"СПбРепетитор" - База репетиторов Санкт-Петербурга

<https://spbrepetitor.ru/> ▼

«СПбРепетитор» - это сообщество преподавателей и педагогов, профессионально занимающихся репетиторством в Петербурге. Наша задача состоит в том, что бы взаимодействие **репетиторов** и учеников было максимально простым и удобным. Если Вам нужен **репетитор**, то вы можете зайти на ...

«Ваш репетитор» — Санкт-Петербург, Ленинградская область ...

spb.repetitors.info/ ▼

Мы — профессиональное сообщество частных **репетиторов**, объединяющее более 220 тысяч преподавателей. На сайте вы можете ознакомиться с анкетами преподавателей и выбрать тех, кто вам подходит, или спросить совета, и мы вам порекомендуем оптимальные варианты. Наши консультации и ...

Ассоциация репетиторов Санкт-Петербурга

<https://spb.repetit.ru/> ▼

База данных **репетиторов** г. Санкт-Петербурга: **репетиторы** и преподаватели иностранных языков, математики, русского языка в Москве. **Репетиторы** по физике, химии, информатике и многие другие.

[Лучшие репетиторы в Санкт-Петербурге с ценами, рейтингом и ...](#)

Алгоритм

Google

маск


Войти

Все Новости Картинки Видео Карты Ещё Настройки Инструменты

Результатов: примерно 4 870 000 (0,58 сек.)


Маск, Илон — Википедия
https://ru.wikipedia.org/wiki/Маск,_Илон
Илон Рив Маск (англ. Elon Reeve Musk /iːlɒn ˈmʌsk/; род. 28 июня 1971 года, Претория, ЮАР) — канадско-американский инженер, предприниматель, филантроп, изобретатель и инвестор; миллиардер. Основатель компании PayPal; основатель, совладелец, генеральный директор и главный инженер ...
Хёрд, Эмбер · Tesla · SpaceX · Neuralink

Главные новости




Истории Бывший рекрутер Tesla рассказал о том, как Илон Маск набирает...

Rusbase
2 дня назад



Илон Маск запустит в космос автомобиль

Gismeteo
2 дня назад



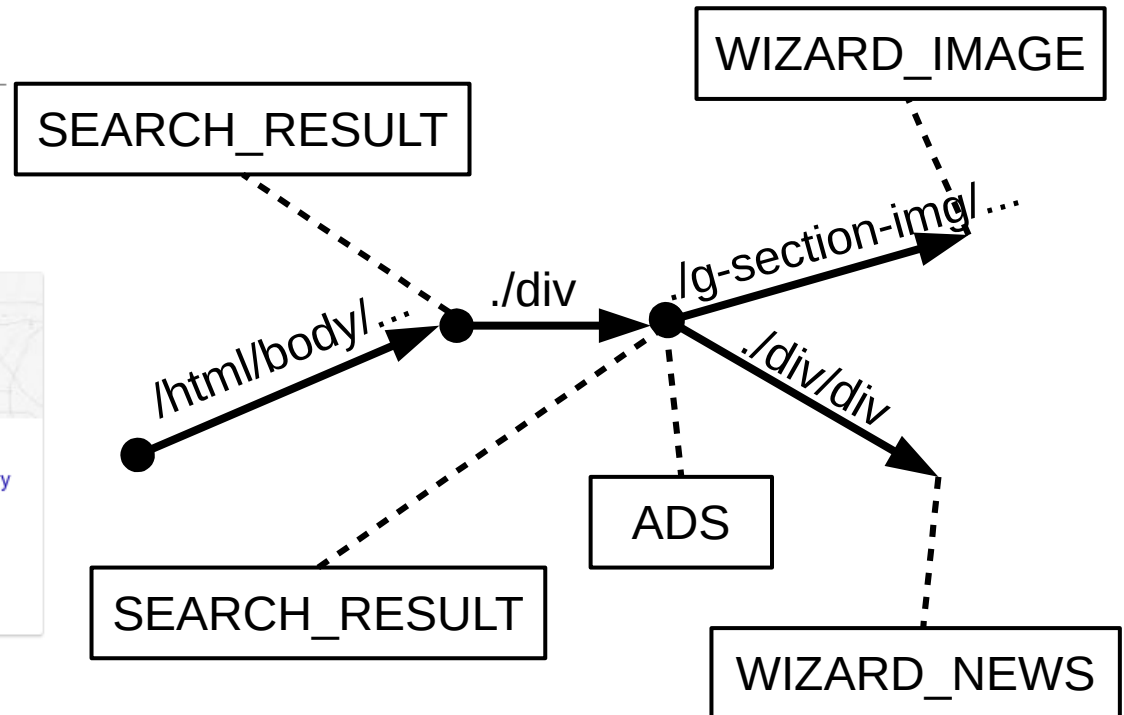
Компания Элона Маска показала карту тоннелей под Лос-Анджелесом

Популярная механика
1 день назад

→ Больше результатов по запросу "маск"

Илон Маск / Elon Musk (Tesla, SpaceX, Neuralink) | ВКонтакте
<https://vk.com/elonmusk>
неофициальная страница* Илона Маска (также Элон Маск, англ. Elon Musk)

Падение Илона Маска началось: СМИ наконец подсчитали его ...
<https://ria.ru/analytics/20171107/1508279232.html>
7 нояб. 2017 г. - Текущие события вокруг, возможно, самого известного бизнесмена в мире Илона Маска приближены к пикову суровых на экономическим пятаками и финансовыми показателями



```
graph LR
    SR1[SEARCH_RESULT] -.-> Node1(( ))
    SR2[SEARCH_RESULT] -.-> Node1
    Node1 -- "/html/body/ ..." --> Node2(( ))
    Node2 -- ".div" --> Node3(( ))
    Node3 -- ".lg-section-img/ ..." --> WI[WIZARD_IMAGE]
    Node3 -- ".div/div" --> WN[WIZARD_NEWS]
    Node3 -.-> ADS[ADS]
```

The diagram illustrates the HTML structure of the search results page. It shows a tree of nodes connected by arrows. Solid arrows represent the main content flow, while dashed arrows represent side elements like ads. The structure starts with two 'SEARCH_RESULT' boxes pointing to a common node, which then leads to a 'body' node, followed by a 'div' node, and finally to 'WIZARD_IMAGE', 'WIZARD_NEWS', and 'ADS'.



Технологии

- Python
- JavaScript
- XPath
- JSON
- CherryPy
- Chrome Extensions API

Репозиторий

https://github.com/kormyshov/parser_generator

Спасибо за внимание

Вопросы?