

Генератор парсеров

Кормышов Михаил

Руководитель: Фёдор Амосов

CSC, осень 2017

Offline оценка качества поиска



Что парсим

Надежный | https://www.google.ru/search?newwindow=1&ei=z_cpWrHUOIy96ASL3lugBw&q=mask&oq=mask&gs_l=psy-ab.3..0l10.4939.5458.0.594

Google




маск

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 /ˈɪlɒn 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 2 дня назад</p>	 <p>Илон Маск запустит в космос автомобиль</p> <p>Gismeteo 2 дня назад</p>	 <p>Компания Элона Маска показала карту тоннелей под Лос-Анджелесом</p> <p>Популярная механика 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/analitics/20171107/1508279232.html>
7 нояб. 2017 г. - Текущие события вокруг, возможно, самого известного бизнесмена в мире Илона Маска публичны в первую очередь не экономическими данными и финансовыми показателями

"components": [
{
"alignment": "LEFT",
"page_url": "/url?q=https://www.ivi.ru/movies&sa=U&ve"
"py/object": "parser_result.Component",
"snippet": "Вы любите смотреть фильмы онлайн и проводи"
"title": "Смотреть фильмы онлайн бесплатно в хорошем"
"type": "SEARCH_RESULT",
"view_url": "https://www.ivi.ru/movies"
},
{
"alignment": "LEFT",
"page_url": "/url?q=https://www.ivi.ru/new/movie-new&
"py/object": "parser_result.Component",
"snippet": "Фильмы-новинки 2016-2017 годов вы можете"
"title": "Фильмы-новинки 2016-2017 годов – смотреть он"
"type": "SEARCH_RESULT",
"view_url": "https://www.ivi.ru/new/movie-new"
},
{
"alignment": "LEFT",
"page_url": "/search?q=%D1%81%D0%BC%D0%BE%D1%82%D1%80"
"py/object": "parser_result.Component",
"title": "Новости по запросу смотреть фильмы",
"type": "WIZARD",
"wizard_type": "WIZARD_NEWS"
},
{
"alignment": "LEFT",
"page_url": "/url?q=http://megogo.net/ru/films&sa=U&ve"
"py/object": "parser_result.Component",
"snippet": "Смотреть фильмы, сериалы, мультфильмы онл"
"title": "Фильмы - смотреть онлайн бесплатно в хороше"

Как парсим 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:
```

Как хотим

The image shows a Google search results page for the query "маск". The search bar at the top contains the word "маск". Below the search bar, the results are categorized by "Все" (All), "Новости" (News), "Картинки" (Images), "Видео" (Video), "Карты" (Maps), "Ещё" (More), "Настройки" (Settings), and "Инструменты" (Tools). The first result is a Wikipedia entry titled "Маск, Илон — Википедия" with the URL "https://ru.wikipedia.org/wiki/Маск,_Илон". The description of the result states: "Илон Рив Маск (англ. Elon Reeve Musk /ˈɪlɒn 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 ...". Below the main result, there is a section titled "Главные новости" (Main news) with three items: "Истории Бывший рекрутер Tesla рассказал о том, как Илон Маск набирает..." from Rusbase, "Илон Маск запустит в космос автомобиль" from Gismeteo, and "Компания Элона Маска показала карту тоннелей под Лос-Анджелесом" from Популярная механика. At the bottom, there are two more search results: "Илон Маск / Elon Musk (Tesla, SpaceX, Neuralink) | ВКонтакте" and "Падение Илона Маска началось: СМИ наконец подсчитали его ...". Annotations with colored arrows point from specific parts of the page to two boxes on the right. The "SEARCH_RESULT" box has labels for "Title" (red), "URL" (green), and "Description" (blue). The "WIZARD_NEWS" box has a label for "URL" (green). The arrows indicate that the Wikipedia result is used for the "SEARCH_RESULT" labels, while the news items and the bottom results are used for the "WIZARD_NEWS" label.

Google

маск

Надежный | https://www.google.ru/search?newwindow=1&ei=z_cpWrHUOIy96ASL3lugBw&q=маск&oq=маск&gs_l=psy-ab.3..0l10.4939.5458.0.594

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 /ˈɪlɒn 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 добавились в первые ряды, не экономическими делами и финансовыми показателями

SEARCH_RESULT
Title
URL
Description

WIZARD_NEWS
URL

Аналоги

ScraperLab



Алгоритм

Надежный | https://www.google.ru/search?newwindow=1&ei=z_cpWrHUOIy96ASL3lugBw&q=mask&oq=mask&gs_l=psy-ab.3..0l10.4939.5458.0.594

Google

mask




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 /ˈɪlɒn 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](#)

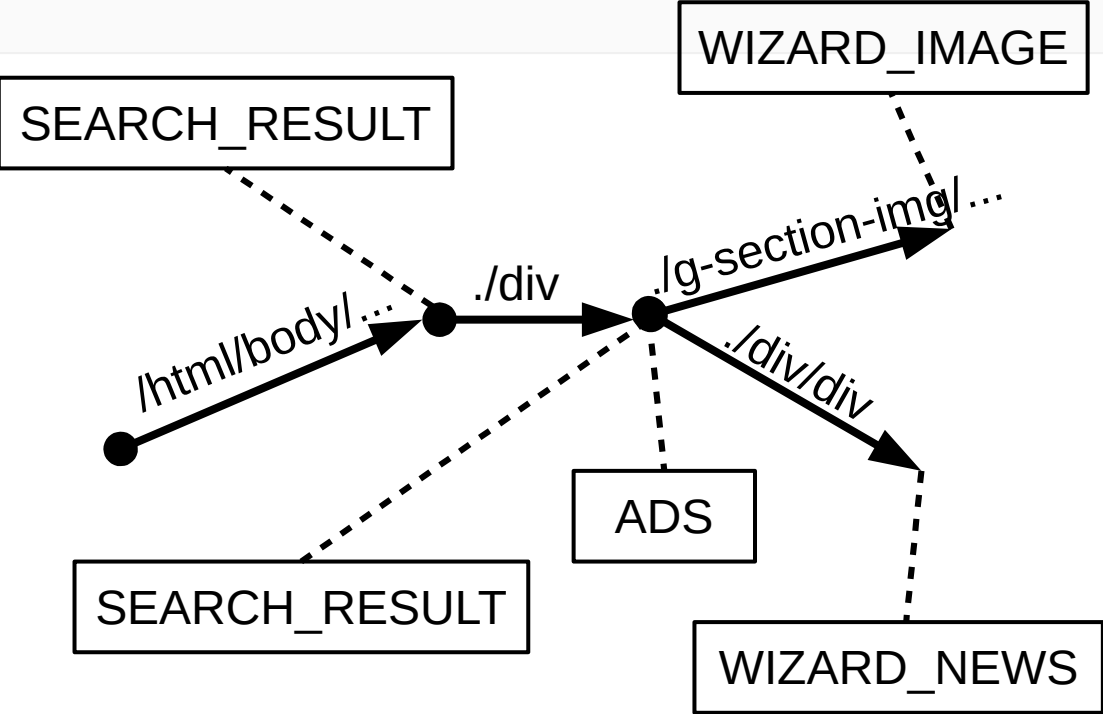
Главные новости

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

[Больше результатов по запросу "маск"](#)

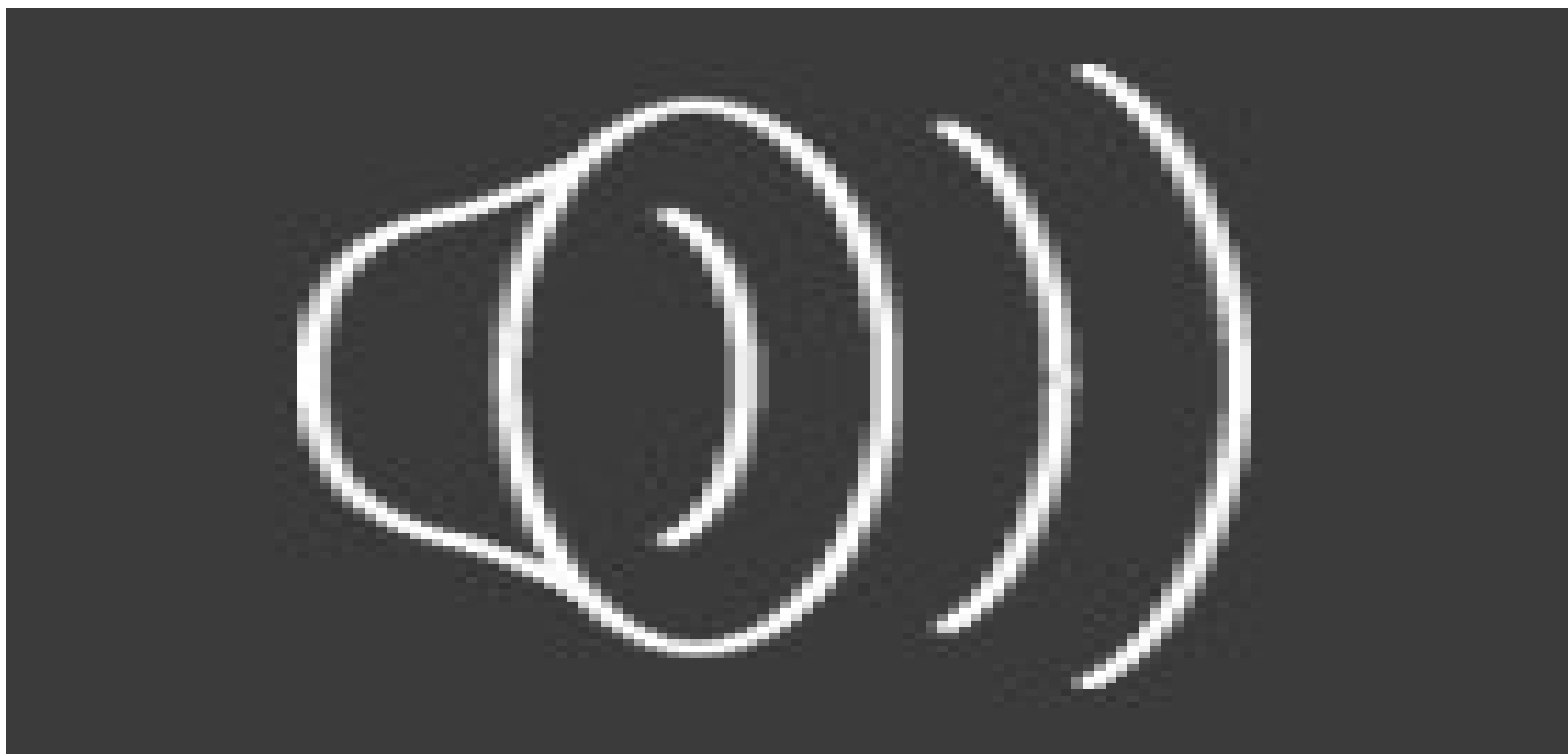
Илон Маск / Elon Musk (Tesla, SpaceX, Neuralink) | ВКонтакте
<https://vk.com/elonmusk>
неофициальная страница* Илона Маска (также Элон Маск, англ. Elon Musk)

Падение Илона Маска началось: СМИ наконец подсчитали его ...
<https://ria.ru/analitics/20171107/1508279232.html>
7 нояб. 2017 г. - Текущие события вокруг, возможно, самого известного бизнесмена в мире Илона Маска публикуются в первую очередь на экономических платформах и финансовыми показателями



```
graph LR
    SR1[SEARCH_RESULT] -.-> B1(( ))
    B1 -- "/html/body/" --> B2(( ))
    B2 -- "/div" --> B3(( ))
    B3 -- "/div/div" --> WZNEWS[WIZARD_NEWS]
    B3 -.-> ADS[ADS]
    B3 -- "/img-section-img/" --> WZIMG[WIZARD_IMAGE]
    SR2[SEARCH_RESULT] -.-> B3
```

The diagram illustrates the search algorithm's path through the HTML structure. It starts with two **SEARCH_RESULT** boxes. The first one points to a node, which then points to another node via the path `/html/body/`. This second node points to a third node via `/div`. From this third node, the algorithm branches: one path goes to **WIZARD_IMAGE** via `/img-section-img/`, another goes to **WIZARD_NEWS** via `/div/div`, and a third goes to **ADS**. A second **SEARCH_RESULT** box also points directly to the third node.



Технологии

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

Спасибо за внимание

Вопросы?