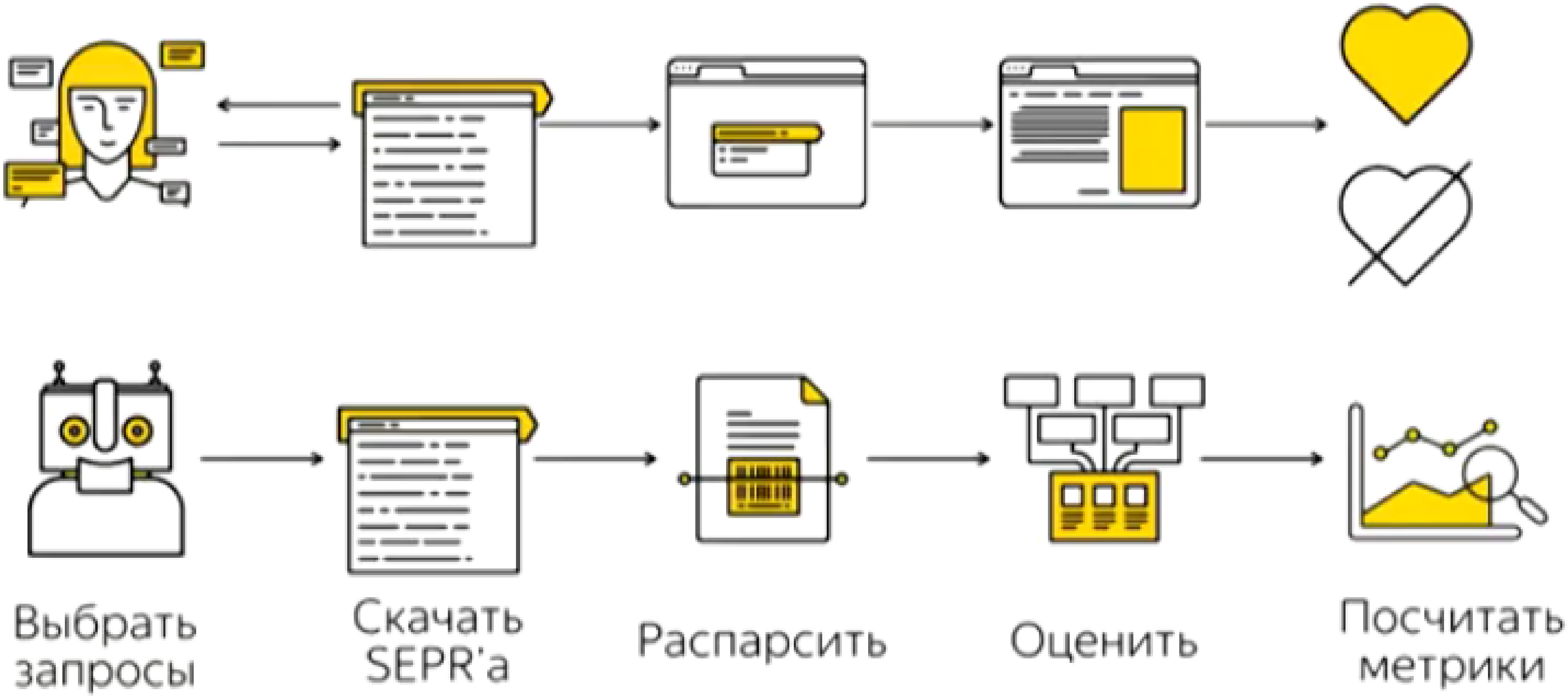
Генератор парсеров

Кормышов Михаил

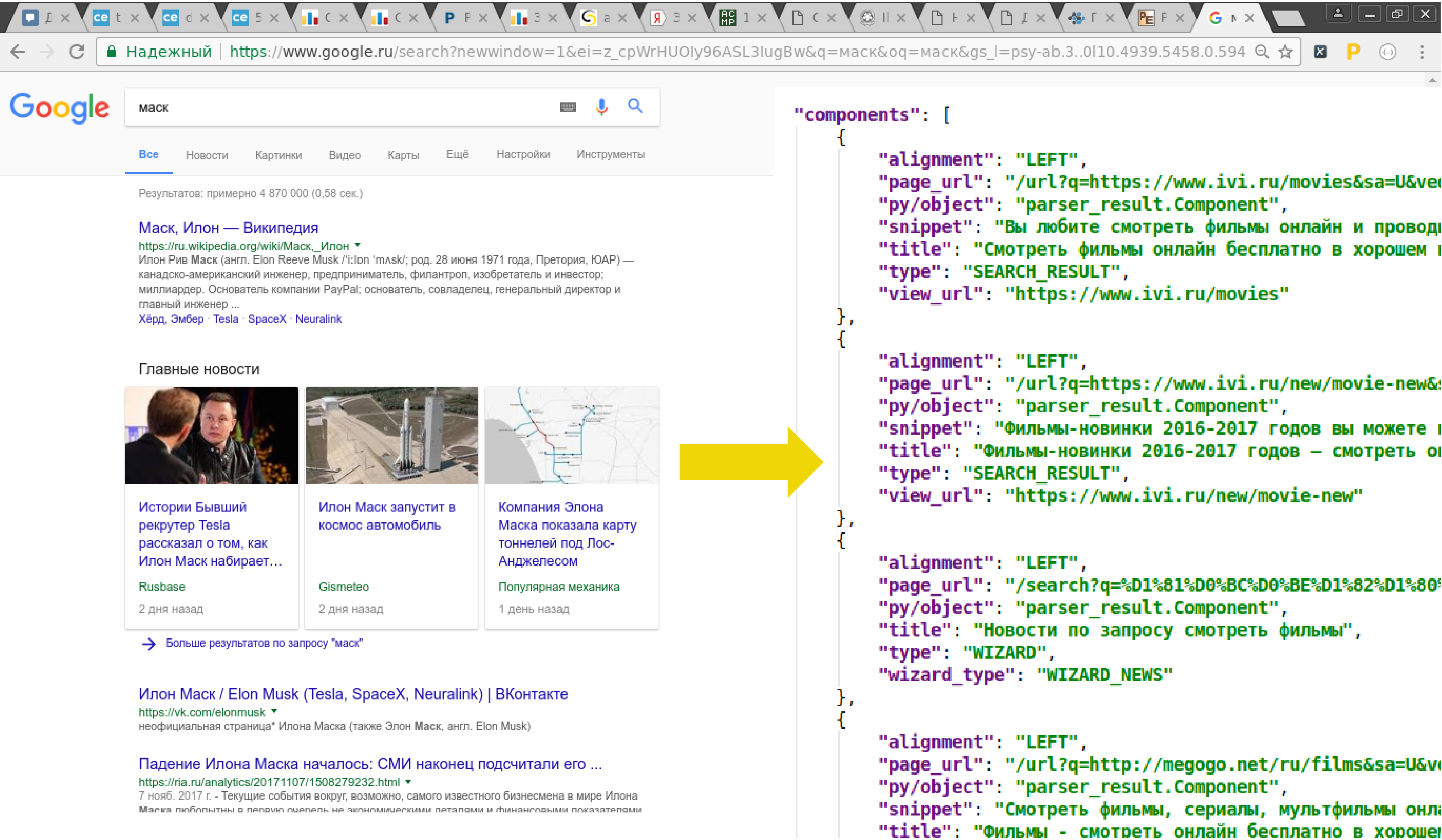
Руководитель: Фёдор Амосов

CSC, осень 2017

Offline оценка качества поиска



Что парсим



Надежный
https://www.google.ru/search?newwindow=1&ei=z_cpWrHUOIy96ASL3IugBw&q=маск&oq=маск&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 /ˈiːlɒn 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

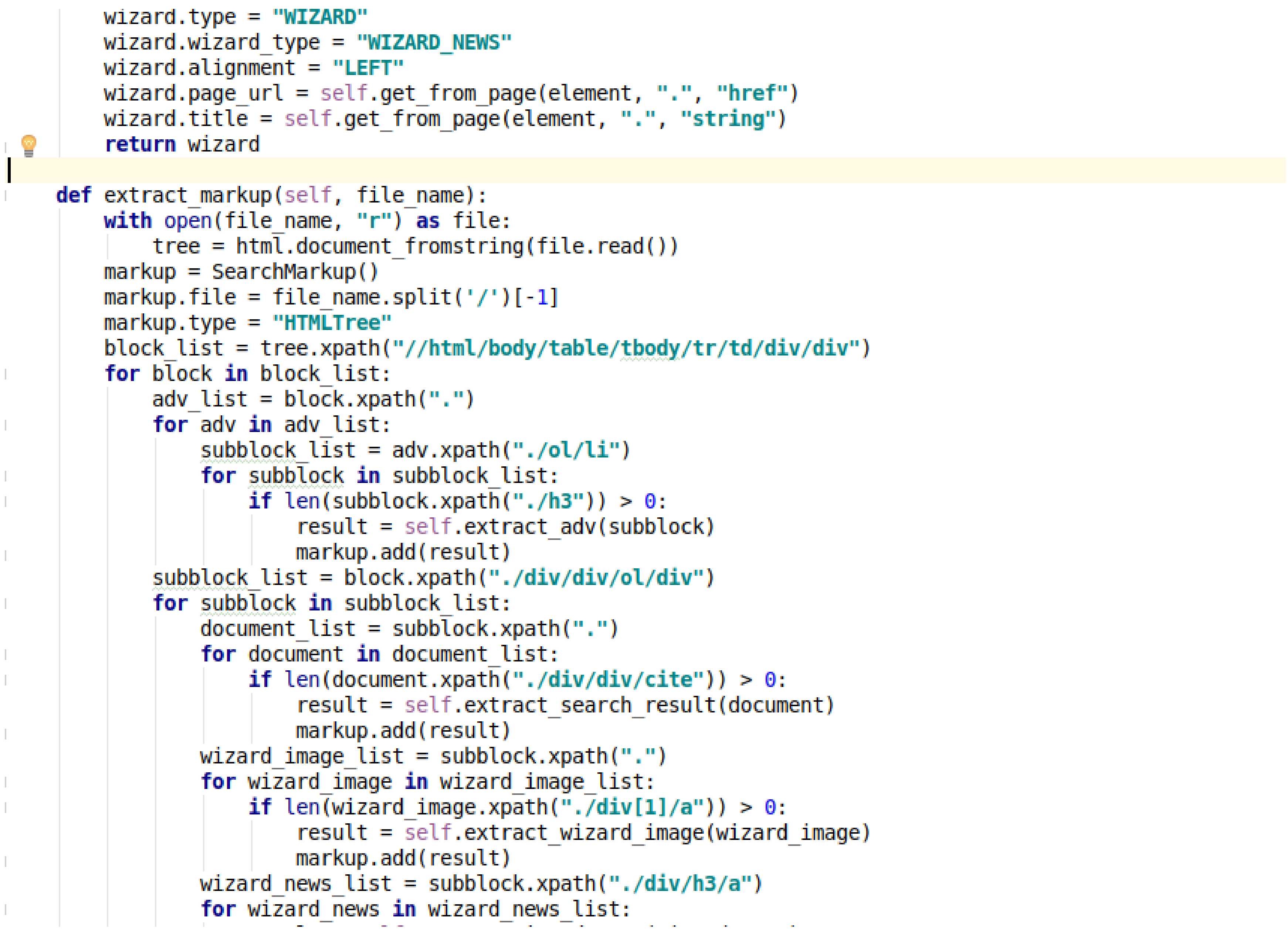
Больше результатов по запросу "маск"

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** любопытны в первую очередь не экономическими деталями и финансовыми показателями

```
"components": [
    {
        "alignment": "LEFT",
        "page_url": "/url?q=https://www.ivi.ru/movies&sa=U&ved
        "py/object": "parser_result.Component",
        "snippet": "Вы любите смотреть фильмы онлайн и проводи
        "title": "Смотреть фильмы онлайн бесплатно в хорошем к
        "type": "SEARCH_RESULT",
        "view_url": "https://www.ivi.ru/movies"
    },
    {
        "alignment": "LEFT",
        "page_url": "/url?q=https://www.ivi.ru/new/movie-new&s
        "py/object": "parser_result.Component",
        "snippet": "Фильмы-новинки 2016-2017 годов вы можете п
        "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 онла
        "title": "Фильмы - смотреть онлайн бесплатно в хорошем
```

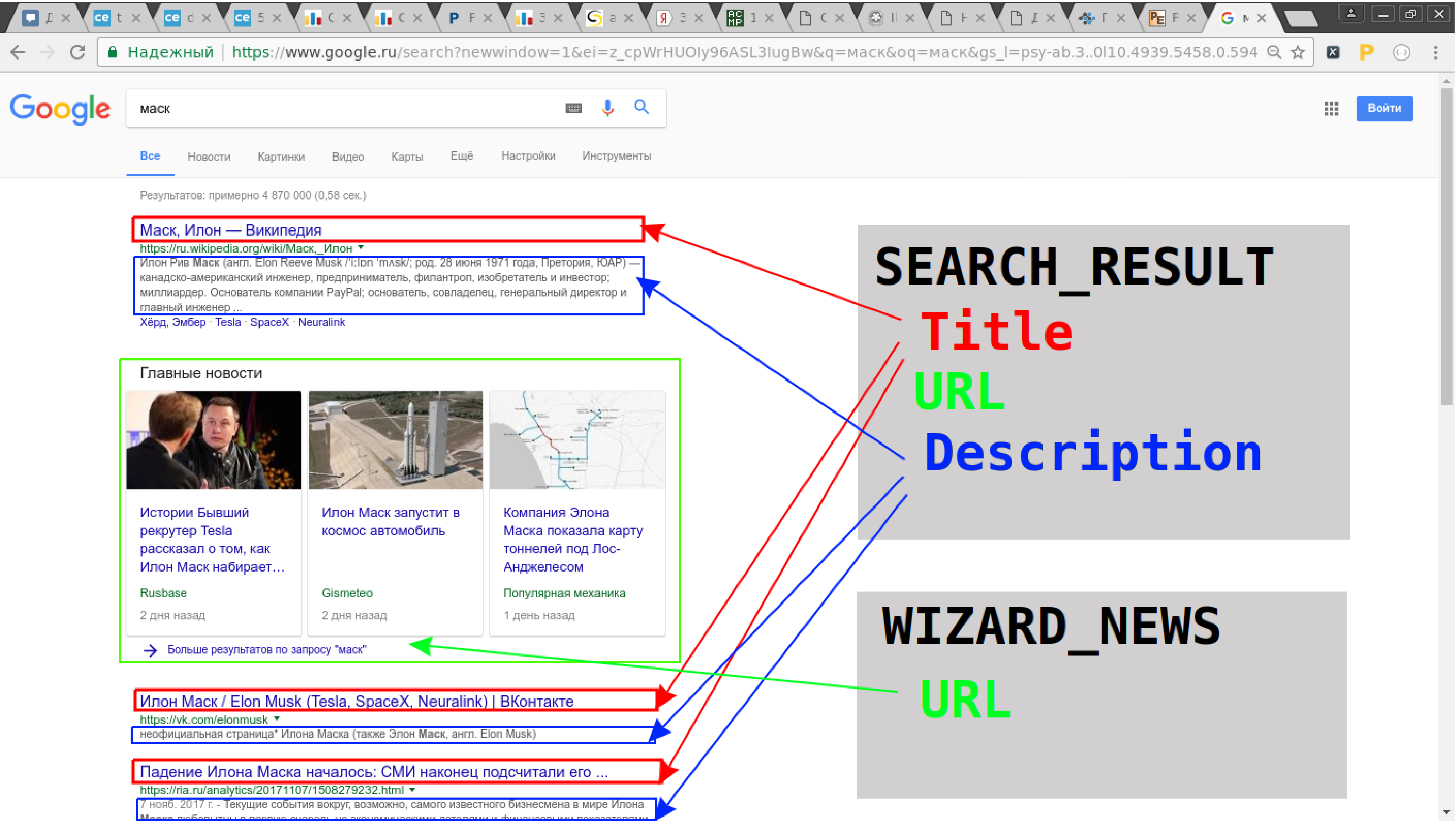
Как парсим сейчас



```
        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


Надежный
https://www.google.ru/search?newwindow=1&ei=z_cpWrHUOIy96ASL3IugBw&q=маск&oq=маск&gs_l=psy-ab.3..0l10.4939.5458.0.594
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 /ˈiːlɒn 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 любопытны в первую очередь не экономическими деталями и финансовыми показателями

SEARCH_RESULT
Title
URL
Description

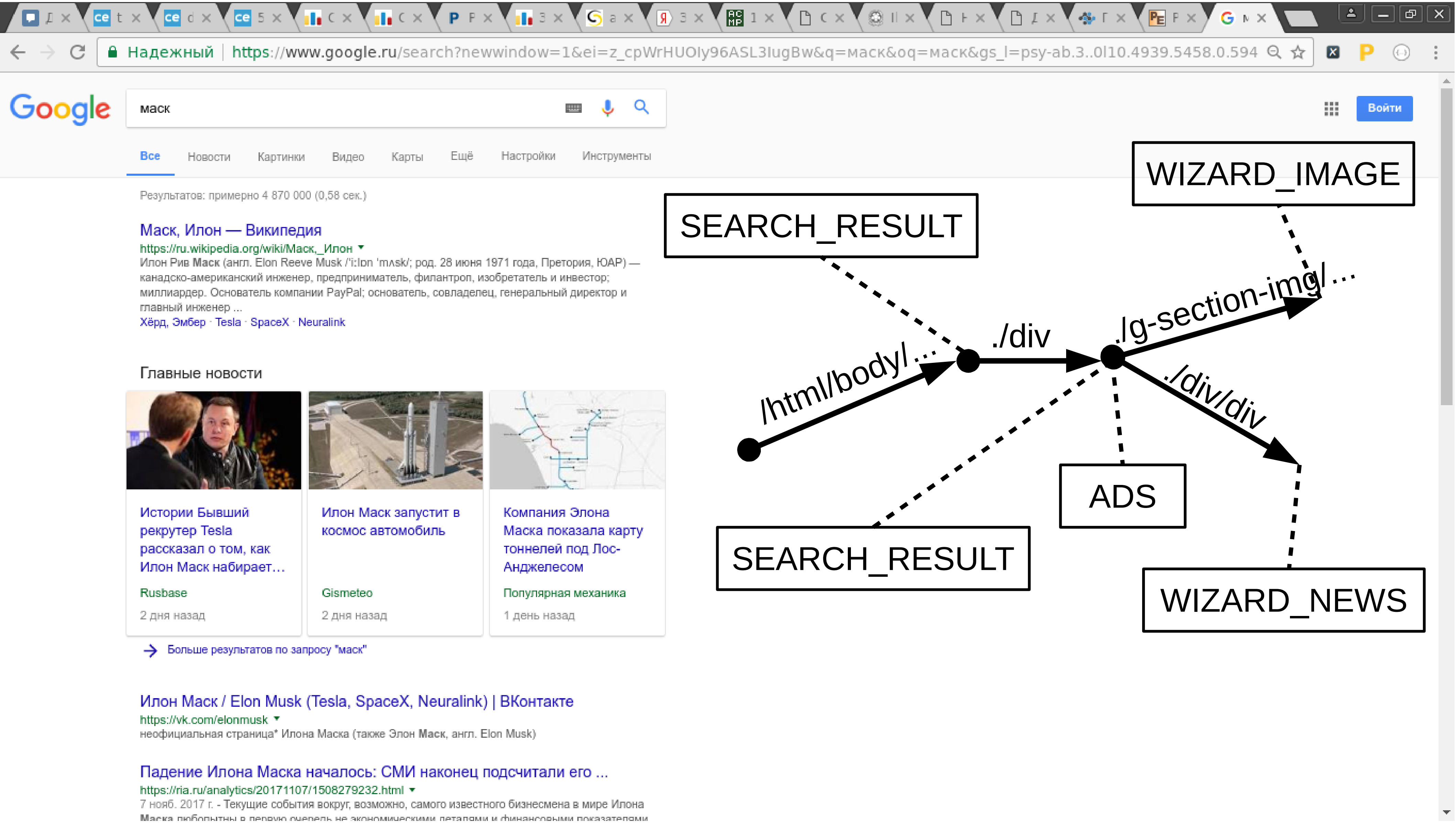
WIZARD_NEWS
URL

Аналоги

ScraperLab

A-PARSER

Алгоритм



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 /ˈiːlɒn 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

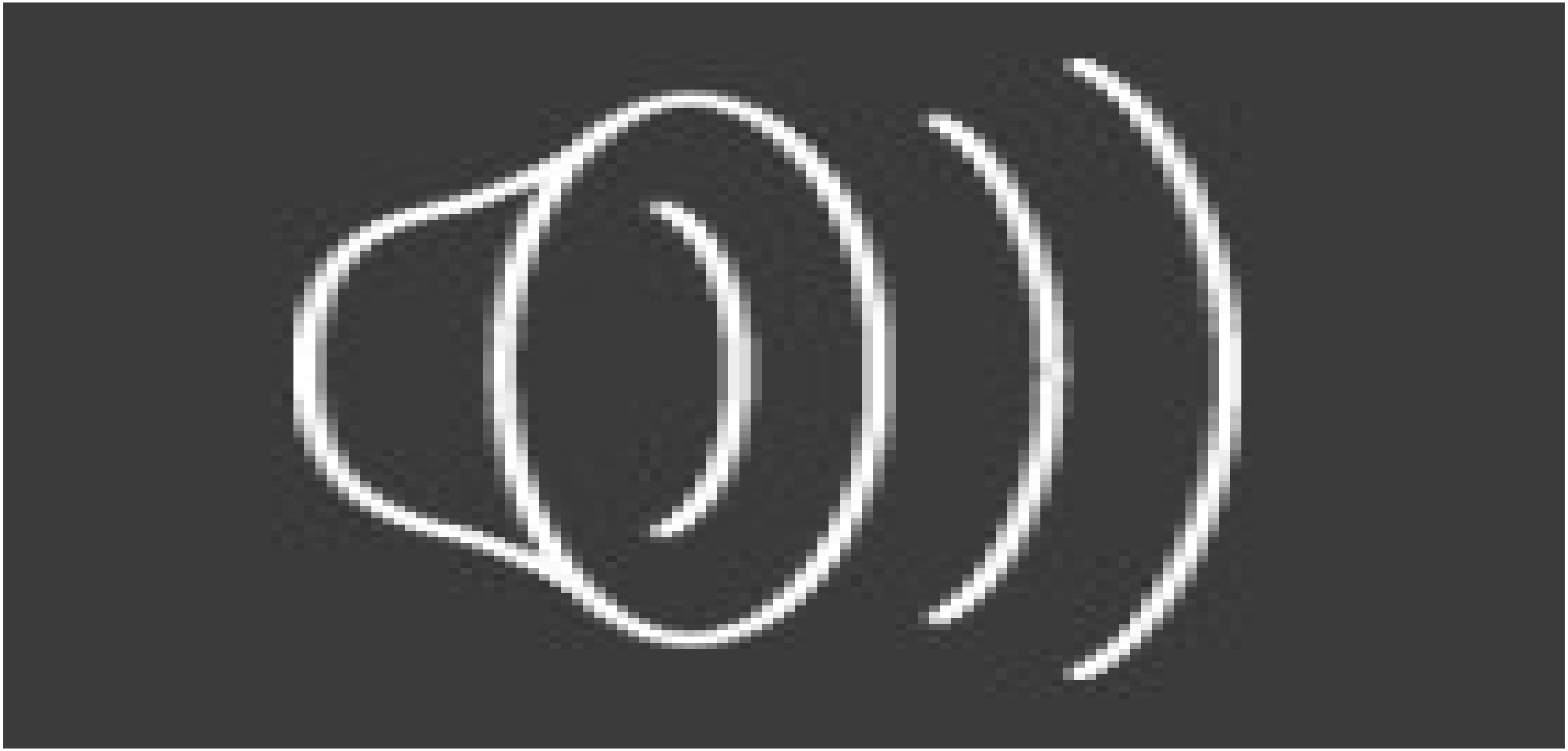
Илон Маск запустит в космос автомобиль
Gismeteo
2 дня назад

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** любопытны в первую очередь не экономическими деталями и финансовыми показателями



Технологии

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

Спасибо за внимание

Вопросы?