

BiByte Butterfly 3.0

Beta 2



Инструкция пользователя

Как искать (быстрый старт)

1. Выберите подходящий Вам профиль поиска, например «Yandex»:

Источник ссылок:

Поиск ссылок Загрузка ссылок из файла

Профиль поиска: Yandex 2011-12-05 [↔] [...]

2. Введите ключевые фразы через запятые или с новой строки:

Ключевые слова:

Ключевые слова Загрузка ключевых слов из файла

Музыка mp3, Скачать музыку, mp3, фильмы, скачать фильмы

3. Выберите подходящий Вам профиль фильтра (чтобы отфильтровать найденные результаты):

Фильтр:

Профиль фильтра: Отбор Drupal 2011-12-07 [↔] [...]

4. Укажите куда сохранять найденные сайты:

Результаты:

Сохранить как: D:/results.txt [...]

5. Нажмите кнопку:

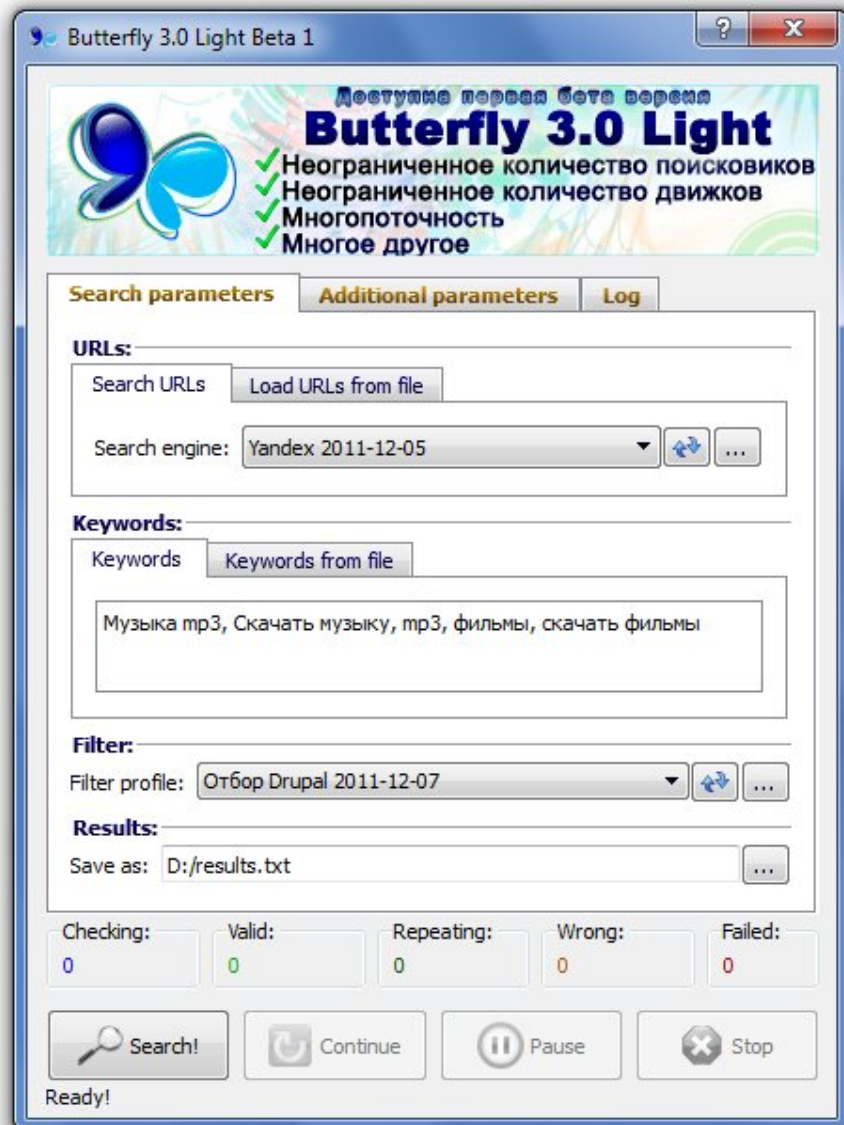


Программа начнет поиск. Все найденные результаты будут сохраняться в указанный файл.

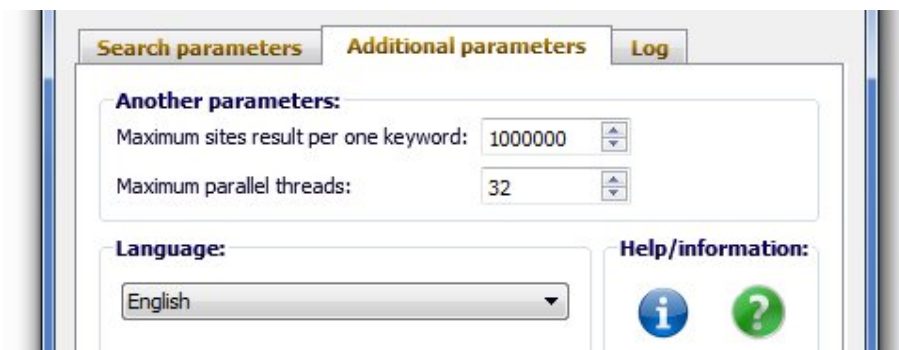
Описание программы

Программа предназначена для парсинга результатов поиска поисковиков, каталогов, страниц сайтов с установленным произвольным отбором ссылок.

После запуска программы открывается главное окно:

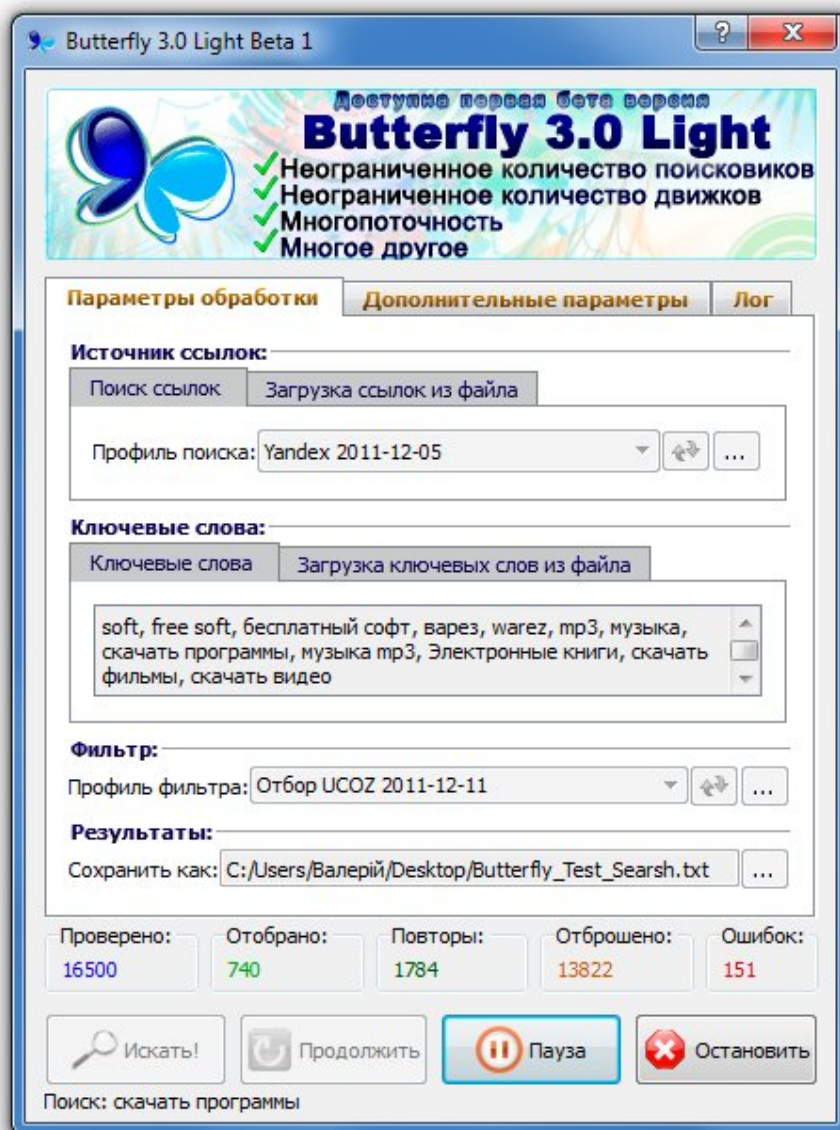


Основная часть окна разделена на три вкладки. Сразу можно перейти на вкладку «**Additional parameters**» и переключить язык интерфейса:



После чего следует перезагрузить программу.

Главное окно в русском интерфейсе выглядит следующим образом:



На вкладке «Параметры обработки» заполняются все необходимые данные для поиска.

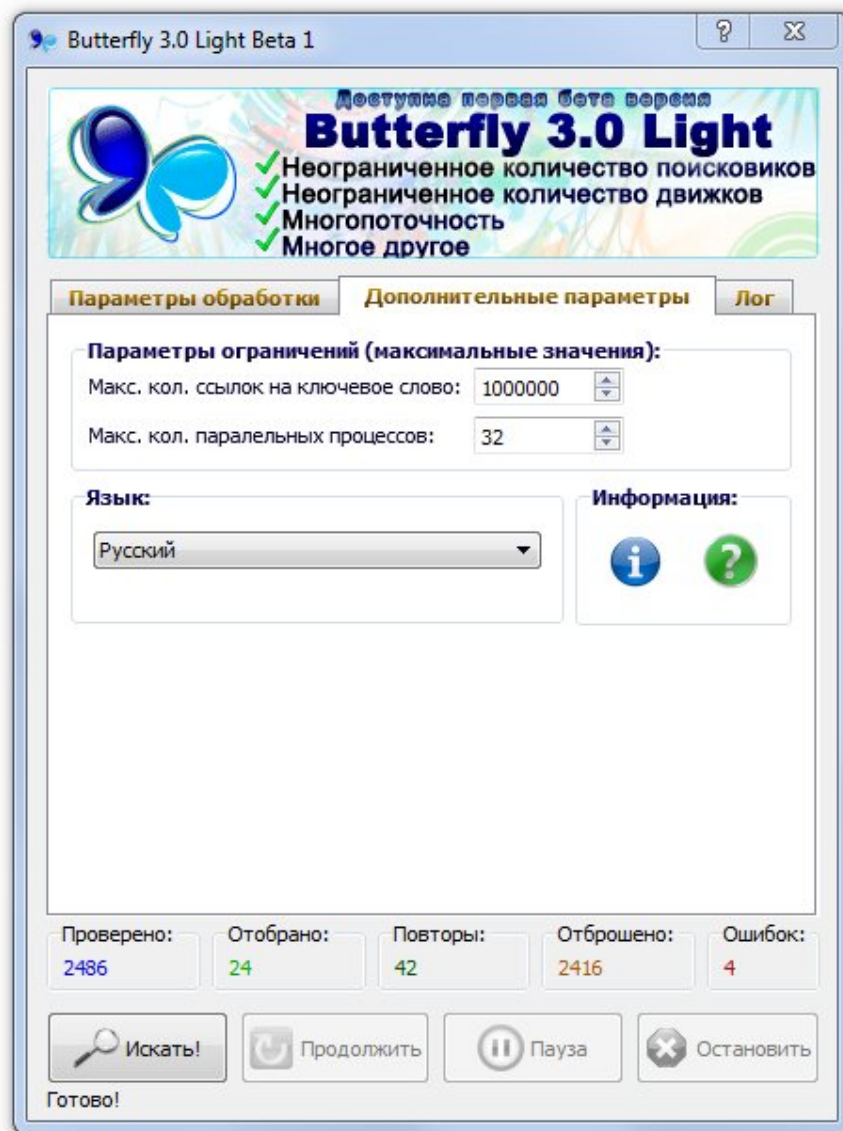
Источник ссылок – источник, откуда программа будет получать ссылки для проверки. Это либо результаты поисковика, либо указывается файл с готовыми ссылками. Если источником должен быть поисковик – выбирается соответствующий профиль поиска.

В поле **Ключевые слова** вводятся ключевые фразы, разделенные запятыми, либо с новой строки. Так же ключевые слова можно загрузить из файла. Для этого нужно перейти на вкладку «**Загрузка ключевых слов из файла**» и указать соответствующий файл.

В поле **Фильтр** указывается дополнительный фильтр, который будет применяться для проверки каждого сайта, полученного из источника. Чтобы не использовать фильтр, нужно в списке выбора «**Профиль фильтра**» выбрать «**Без отбора**».

В поле «**Результаты**» выбирается файл, куда будут сохраняться найденные и отфильтрованные сайты. Как только программа нашла и отфильтровала очередной сайт, он тут же записывается в файл результатов, поэтому можно не волноваться, что при аварийном закрытии программы полученные результаты будут утеряны.

На вкладке «**Дополнительные параметры**» расположены некоторые другие параметры, которые используются реже:



Макс. кол. ссылок на ключевое слово – максимальное количество ссылок, которое будет проверяться программой для одной ключевой фразы. После поиска, когда программа проверила число результатов, равное этому параметру, программа передает поисковику следующее ключевое слово, которое тоже проверяет не более максимального количества. Если количество ссылок не достигает максимального, а поисковик больше не имеет результатов – программа переходит к поиску по следующей ключевой фразе.

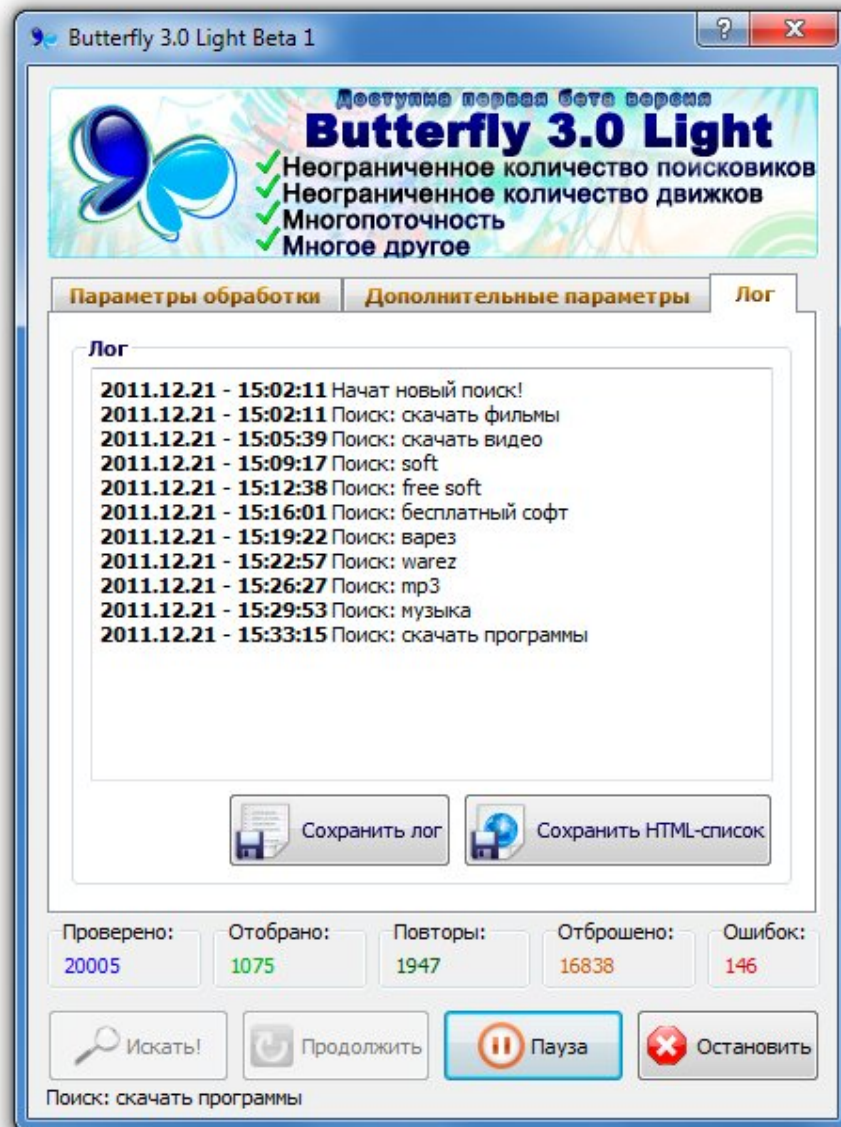
Макс. кол. паралельных процессов – максимально допустимое количество параллельных потоков проверки сайтов. Чем больше значение – тем быстрее программа будет проверять сайты, но тем

большая нагрузка на ресурсы ПК. Для ПК Intel Core2Duo 3Ghz, 2Gb RAM не рекомендуется устанавливать это значение более 400 (В линейке **Butterfly Light** это значение ограничено до 32).

Язык – Переключение языка интерфейса. Для того чтобы изменения вступили в силу, необходимо перезапустить программу.

Информация – панель информации. Здесь расположены кнопки открытия этой справки и информационного окна «Про программу».

На вкладке «**Лог**» ведется лог всех действий и состояний программы:



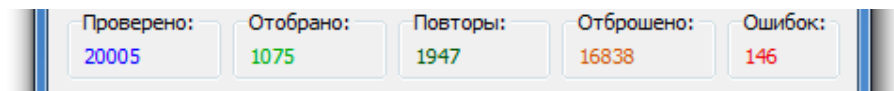
Здесь расположены дополнительные кнопки:

Сохранить лог – текущий лог сохраняется в .txt файл.

Сохранить HTML-список – данная кнопка дает возможность сохранить последнюю страницу результатов поиска, полученную программой от поисковика для дальнейшего анализа работы программы с данным поисковиком. Эта функция полезна при создании профилей поиска и фильтрации.

Нижняя часть программы, расположенная под основной панелью содержит информационное поле, показывающее состояние поиска, а так же панель кнопок управления программой.

Панель информационного поля имеет следующий вид:



На ней отображается количество **проверенных** ссылок, из них:

отобранные – те которые прошли фильтр;

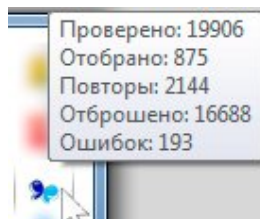
повторы – те которые прошли фильтр, но уже имеются в файле с результатами;

отброшено – те которые не прошли фильтр;

ошибок – количество сайтов, которые были недоступны в момент проверки.

Всегда **Проверено = Отобрано + Повторы + Отброшено + Ошибок**.

Так же эта информация отображается при наведении курсора на значок программы в трее:



Панель управления программой имеет следующие кнопки:

Искать! – Начинает поиск по установленным параметрам;

Пауза – Приостанавливает поиск;



Остановить – Останавливает поиск, сбрасывая текущее состояние программы;

Продолжить – Продолжает поиск после паузы либо аварийной остановки.

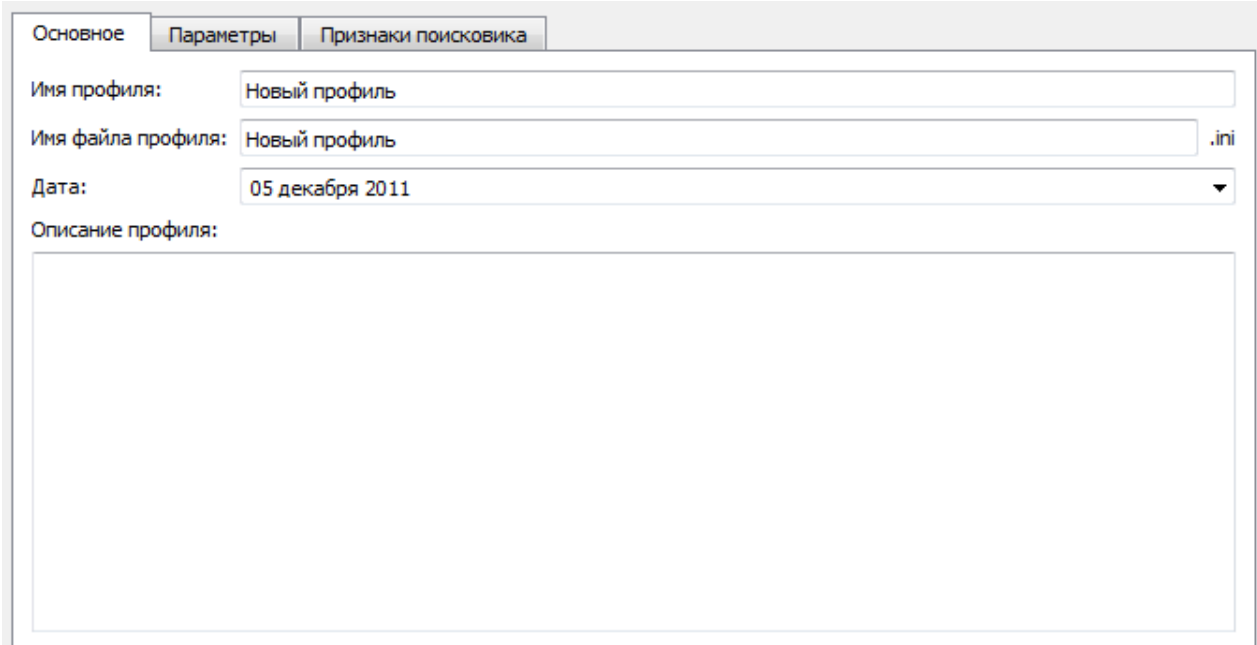
При работе программы, если программа перейдет в аварийное состояние (пропало соединение с интернетом, поисковик не доступен или др.), а так же по окончании поиска, программа переходит в состояние паузы. Чтобы можно было продолжить поиск из текущего места, когда наступят благоприятные условия.

После окончания поиска программа так же переходит в состояние паузы. Необходимо нажать кнопку **«Остановить»**, тогда поля настроек вновь станут доступными, и можно будет начать новый поиск.

Создание и редактирование профилей поиска

Для создания нового профиля нажмите на кнопку . Для того, чтобы создать профиль, скопировав какой-то из существующих, выделите существующий профиль и нажмите кнопку .

При создании нового профиля, он заполняется следующими данными:



The screenshot shows a dialog box with three tabs: 'Основное' (Basic), 'Параметры' (Parameters), and 'Признаки поисковика' (Searcher's signs). The 'Основное' tab is active. It contains the following fields:

- Имя профиля:** Новый профиль
- Имя файла профиля:** Новый профиль .ini
- Дата:** 05 декабря 2011
- Описание профиля:** (empty text area)

На вкладке «**Основное**» необходимо ввести имя профиля, имя файла профиля, дату создания и описание. Дата профиля обычно ставится текущей, чтобы определить дату актуальности профиля. В описание профиля вводится информация о профиле.

На вкладке «**Параметры**» необходимо заполнить данные, которые позволят осуществлять перебор страниц поисковика:

Основное **Параметры** Признаки поисковика

Значения параметров

Адрес поиска:

Начальная страница:

Шаг страницы:

[q] - параметр запроса. Во время поиска он будет заменяться ключевыми фразами.
[p] - параметр страницы результатов. Во время поиска он будет заменяться текущей страницей поисковых результатов.
Пример: `http://host.com/srch?query=[q]&page=[p]`

Поисковый адрес поочередно принимает следующий вид:
`http://host.com/srch?query=Free soft&page=1`
`http://host.com/srch?query=Free soft&page=2`
`http://host.com/srch?query=Free soft&page=3`
`http://host.com/srch?query=Free mp3&page=1`
`http://host.com/srch?query=Free mp3&page=2`

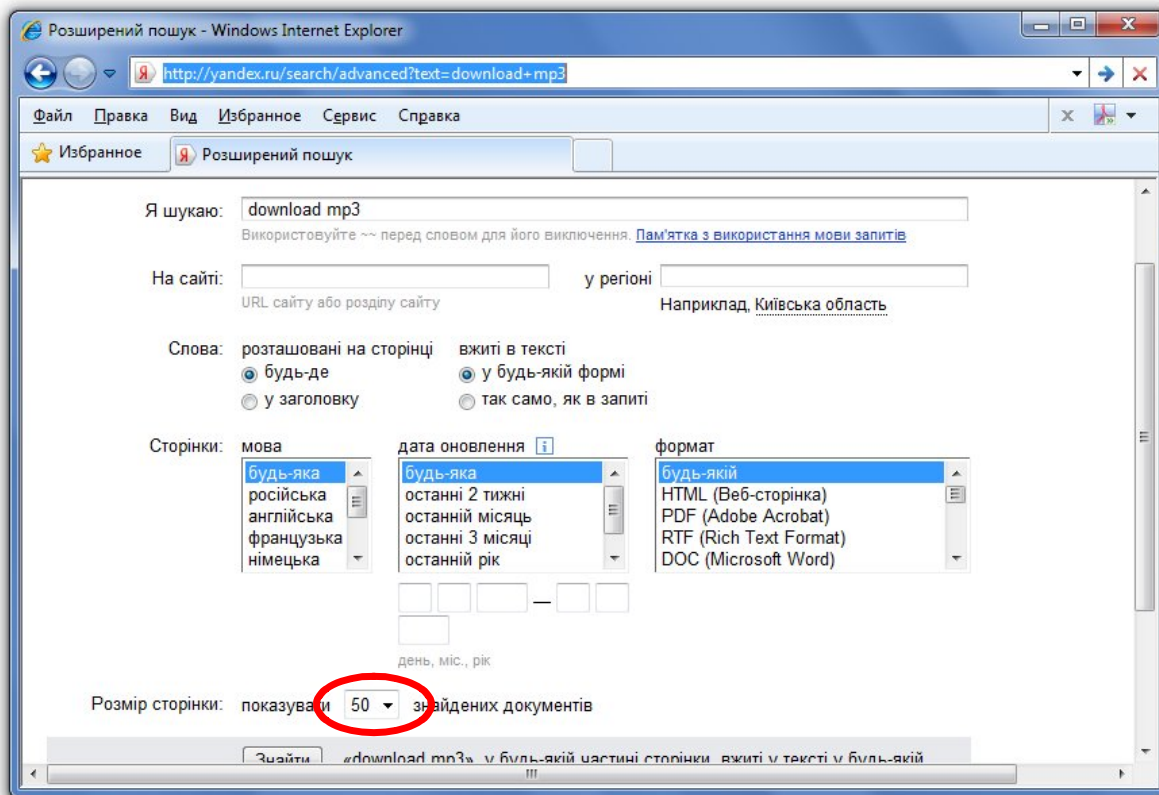
В поле **Адрес поиска** необходимо ввести параметрическую строку адреса поисковика, определив параметры поискового запроса [q] и текущей страницы [p]. Во время поиска вместо параметра [q] будет подставляться реальный поисковый запрос, а вместо [p] – номер страницы, начальная позиция которой определяется значением «Начальная позиция», а шаг значением параметра «Шаг страницы».

Рассмотрим реальный пример заполнения таких параметров на поисковике Yandex.

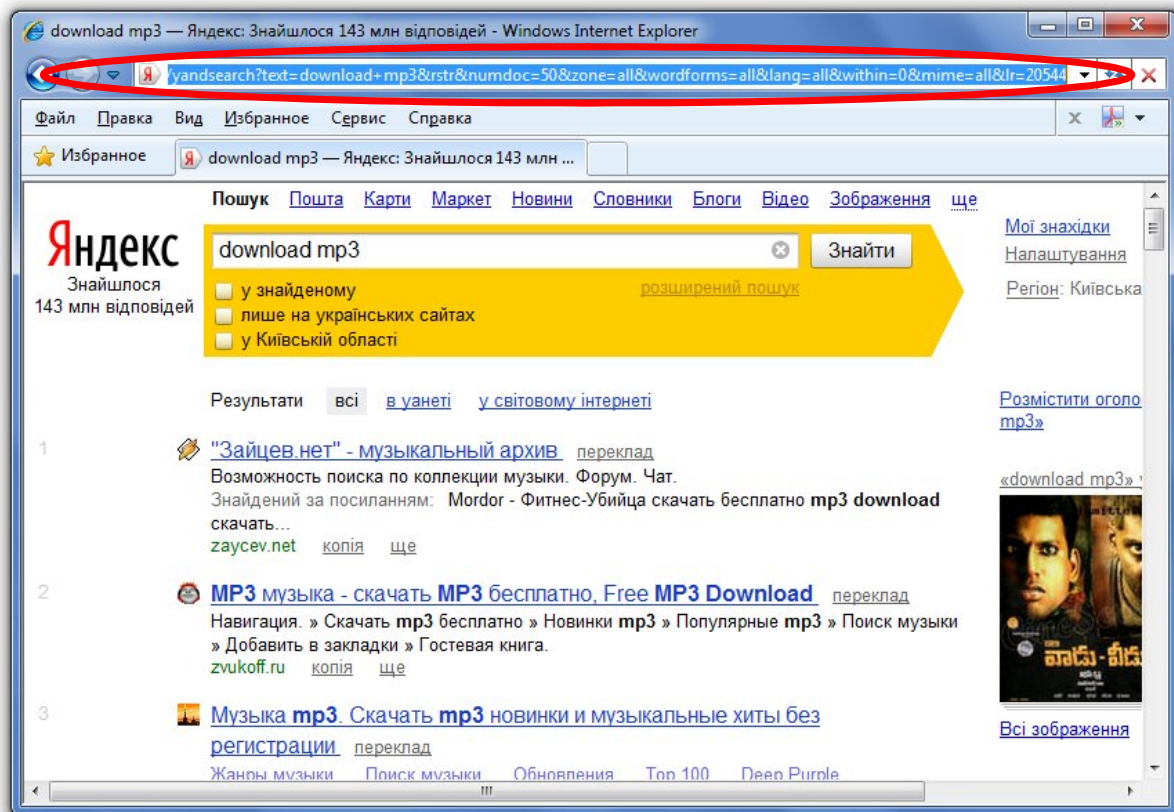
Основные параметры поисковика

Сначала получим параметрический адрес поиска. Для этого достаточно зайти на страницу поиска, ввести поисковый запрос и нажать «Искать». Поисковик выдаст результаты, а в поле адреса сформируется нужная нам строка. Именно так как сейчас ищет поисковик будет искать и программа, поэтому в поисковике можно настроить расширенные параметры поиска, чтобы поисковик выдавал как можно большее количество результатов, или результаты только на английском или другом языке. Таким образом, Вы сможете создать профиль поиска на конкретном языке или из конкретной страны.

Настроив дополнительные параметры, нажмем кнопку поиска:



Поисковик выдал результаты, но кроме того сформировал адресную строку по которой осуществляется поиск с расширенными параметрами:



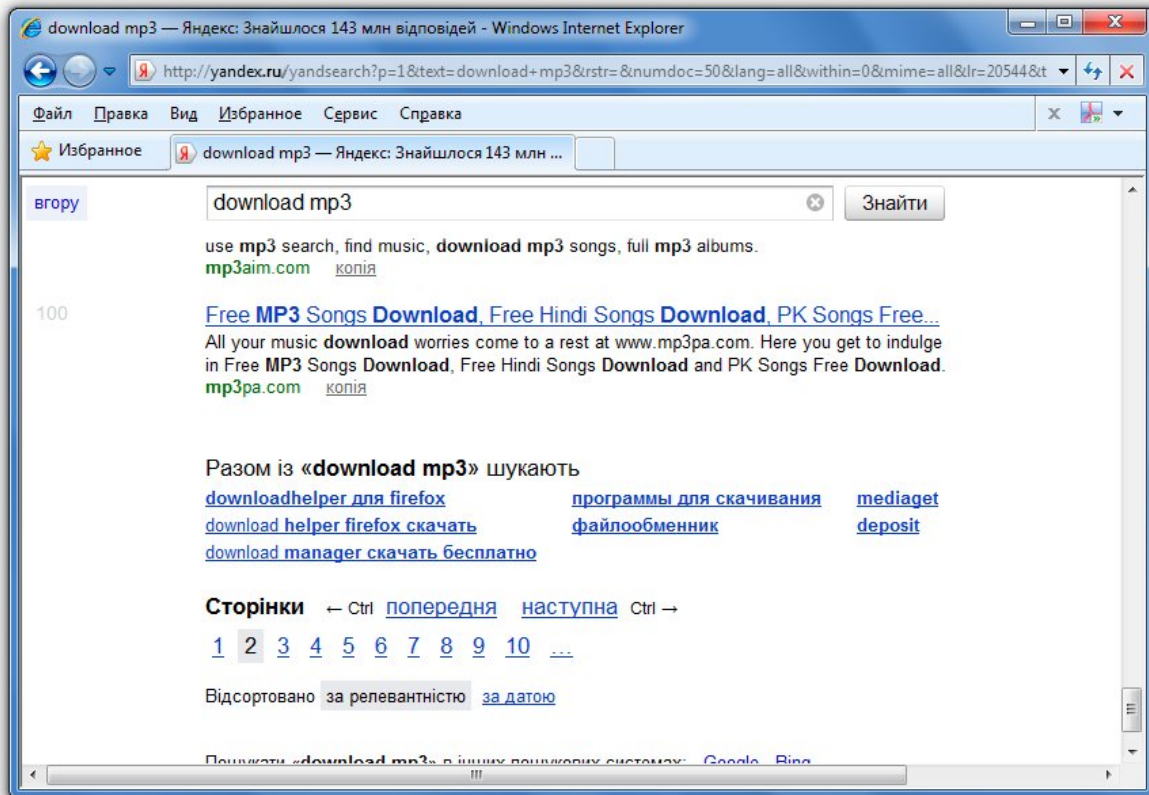
Мы получили адресную строку:

<http://yandex.ru/yandsearch?text=download+mp3&numdoc=50>

В строке мы оставили самые важные параметры – поисковый запрос и параметр вывода результатов по 50 на страницу. Другие параметры на результат поиска особо не влияют. Очевидно, что параметр поискового запроса это «**text**», потому вместо поискового запроса подставим **[q]**:

[http://yandex.ru/yandsearch?text=\[q\]&numdoc=50](http://yandex.ru/yandsearch?text=[q]&numdoc=50)

Однако наш адрес не имеет параметра номера страницы. Если такового нет, нужно просто перейти на следующую страницу поиска и тогда он появится:



Как мы видим, на второй странице появился параметр «**p=1**». Это и есть параметр страницы, потому окончательно адресная строка будет выглядеть так:

[http://yandex.ru/yandsearch?text=\[q\]&numdoc=50&p=\[p\]](http://yandex.ru/yandsearch?text=[q]&numdoc=50&p=[p])

Если мы теперь перейдем на третью страницу, параметр «**p**» примет значение **2**. Как мы видим, нумерация страниц начинается с 0 с шагом = 1. Потому параметр начальной страницы установим в «0», а шаг страницы в «1».

Таким образом, страница параметров будет выглядеть так:

Основное Параметры Признаки поисковика

Значения параметров

Адрес поиска:

Начальная страница:

Шаг страницы:

[q] - параметр запроса. Во время поиска он будет заменяться ключевыми фразами.
[p] - параметр страницы результатов. Во время поиска он будет заменяться текущей страницей поисковых результатов.
Пример: `http://host.com/srch?query=[q]&page=[p]`

Поисковый адрес поочередно принимает следующий вид:
`http://host.com/srch?query=Free soft&page=1`
`http://host.com/srch?query=Free soft&page=2`
`http://host.com/srch?query=Free soft&page=3`
`http://host.com/srch?query=Free mp3&page=1`
`http://host.com/srch?query=Free mp3&page=2`

P.S. Если мы ориентируемся не только на русскоязычных пользователей, лучше использовать не привычный нам домен «**yandex.ru**», а более общий. Для яндекса это «**yandex.net**»

Определение признаков поисковика

Кроме обеспечения перебора страниц поисковика, необходимо обеспечить определение того, что страницы с результатами уже закончились, результаты отсутствуют вовсе или выдана страница, оповещающая Вас о том, что Вы забанены. В первых двух случаях программа просто начинает поиск по следующему ключевому слову. В третьем – приостанавливает поиск.

Основное Параметры Признаки поисковика

Последняя страница: ☐

Нет результатов: ☐

Подозрение в спаме: ☐

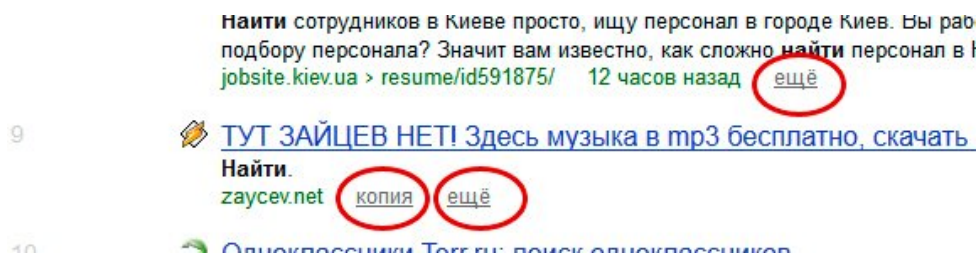
Кеш:

Если некоторые из параметров не заполнять, проверка их не будет осуществляться. Однако если параметр определения последней страницы не указан, то программа будет бесконечно перебирать страницы поисковика по одному ключевому слову. В таком случае Вы заметите, что программа перестала перебирать ссылки, или перебирает по 5-6 в минуту.

Так же само, если не указать параметр определения того, что результатов по ключевому слову нет, если такой случай произойдет, программа будет вести себя как в первом случае.

Параметр «**Подозрение в спаме**» можно заполнять в том случае, если произошел бан (это происходит с немногими поисковиками). Программа будет вести себя, так же как и в первых двух случаях. Для определения причины следует остановить поиск, перейти на вкладку «Лог» и нажать «**Сохранить HTML-список**» - это именно та страничка, которую поисковик пытается парсить в данный момент. Если после сохранения вы откроете ее в браузере и уведете что-то типа вывода капчи и похожих сообщений (или редиректа на страницу с капчей), там можно и найти какой-то признак, по которому программа будет выявлять эту страницу. Так же само можно определить страницу отсутствия результатов и последней страницы.

Часто поисковики вместе с основными ссылками выдают ссылки на кеш, которые бывает полезно отсеять:



Для этого в профиле нужно заполнить признак «Кеш» и ввести туда фрагмент ссылки на кеш. Обычно такие ссылки содержат название поисковика.

Например, ссылка на кеш Яндекса имеет вид:

http://hghltd.yandex.net/yandbtm?fmode=inject&url=http%3A%2F%...

Таким образом, каждая такая ссылка на точно содержит в себе фрагмент «**.yandex.net**». Именно этот фрагмент нужно вставить в поле «**Кеш**», тогда программа не будет учитывать эти ссылки при поиске.

В каждое поле признаков нужно заполнить текст, который содержится только на той странице поисковика, которая является последней в результатах поиска, либо не имеет результатов и установить галочку напротив поля. Или заполнить текст, которого нет только на той странице, а на остальных есть. Галочку тогда нужно снять.

Например, если HTML- код страницы поисковика, которая сообщает, что результатов не найдено, содержит текст «**div class="msg_no_results"**», а все остальные страницы поисковика не содержат этот текст, тогда данный текст можно принять как признак отсутствия результатов. Его следует вписать в поле «Нет результатов» и установить галочку напротив.

Другой пример: если HTML- код любой страницы поисковика содержит текст «**div class="result_list"**», кроме той страницы, которая сообщает что результатов не найдено. Этот текст тоже можно считать признаком отсутствия результатов поиска (в случае если он будет не обнаружен). Тогда его следует вписать в поле «Нет результатов» и снять галочку напротив.

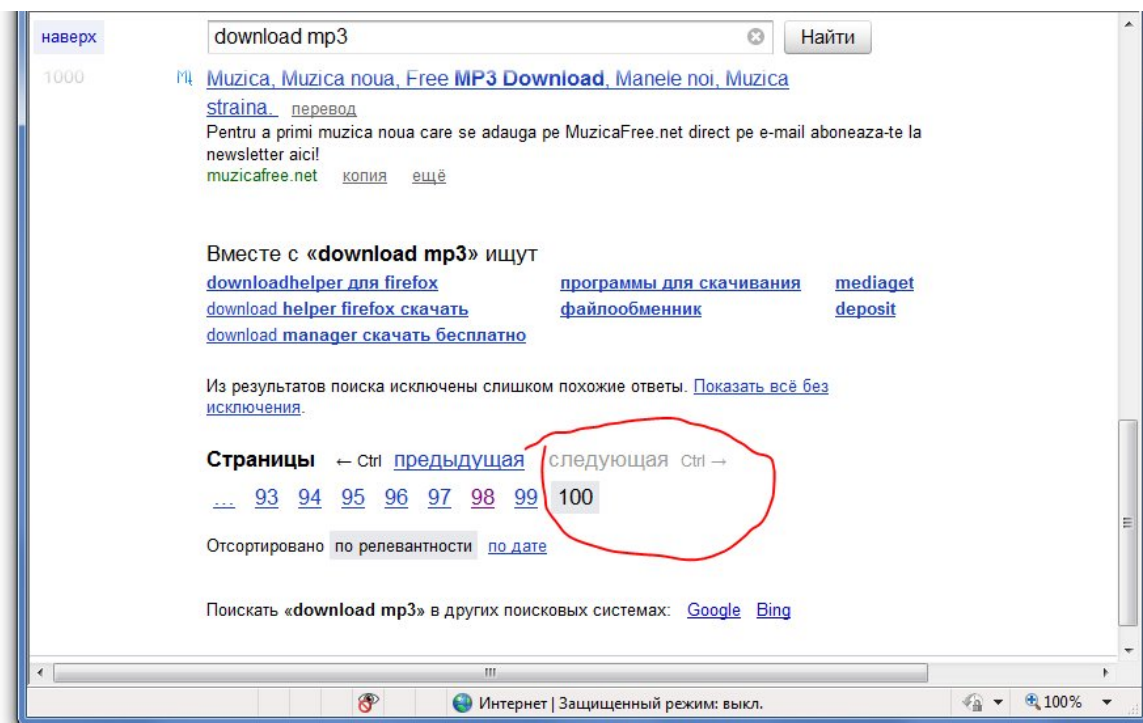
Галочки напротив поля параметров означают, что условие должно выполняться, когда указанный фрагмент текста присутствует на странице (галочка установлена) или отсутствует (галочка снята).

Определение признака окончания результатов поисковика

Найдем признак окончания результатов поисковика для Яндекса.

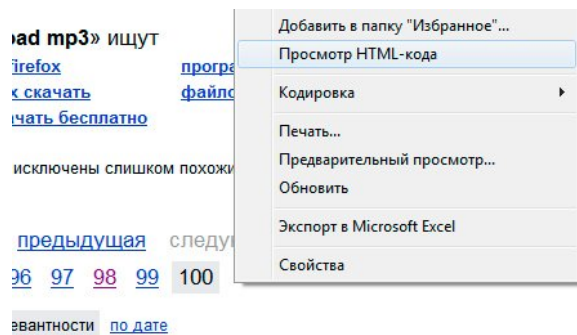
1. Перейдем на страницу поиска, введем поисковый запрос и запустим поиск.

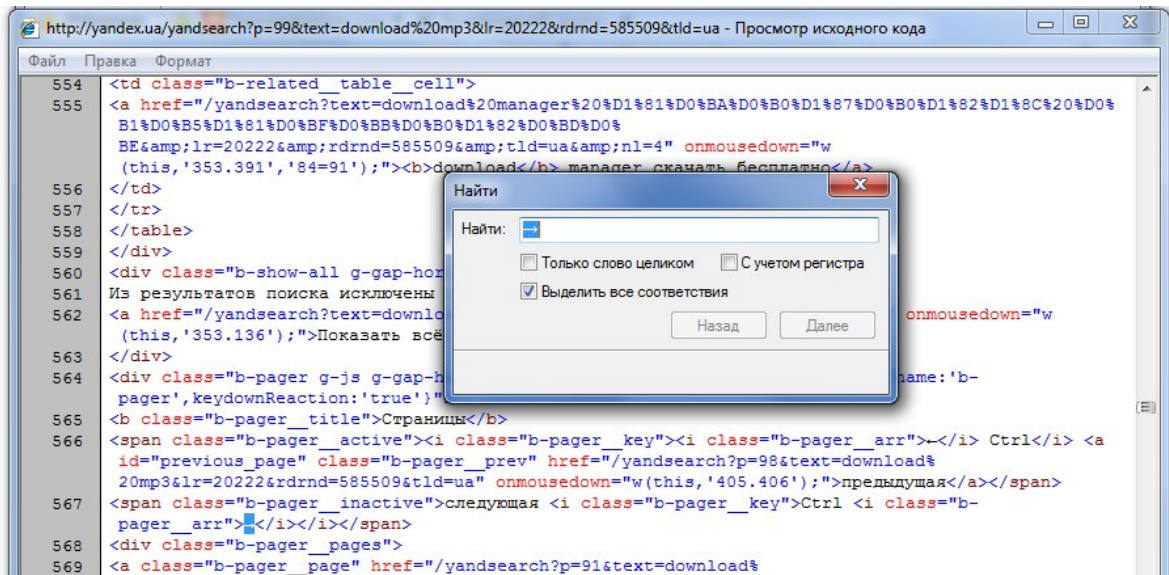
2. Перейдем на последнюю страницу результатов поиска:



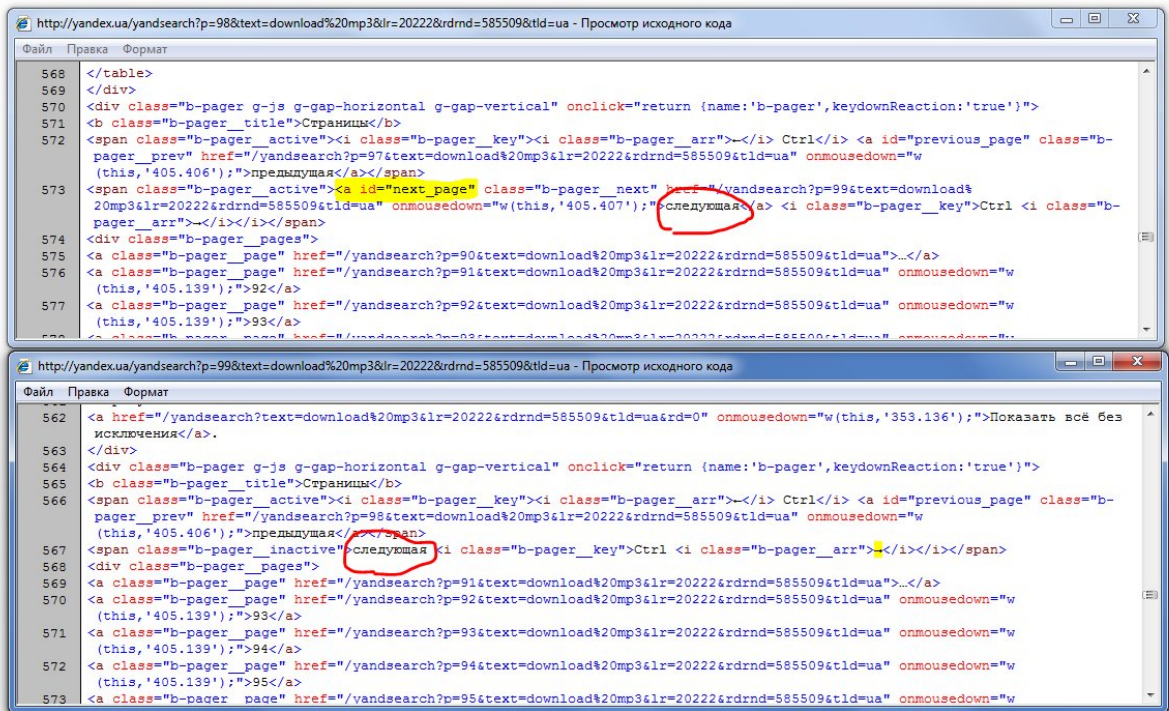
Как видим ссылки «[следующая](#)» и «[Ctrl ->](#)» не активны, а страницы 101 вовсе нет (то есть, нет справа ссылки от текущей страницы 100). Они становятся недоступными только на последней странице поиска (и возможно на странице с 2-3-мя результатами). Именно в коде их описания содержится текст, которого нет на других страницах, и который может быть признаком последней страницы.

3. Зайдем в код этих ссылок:





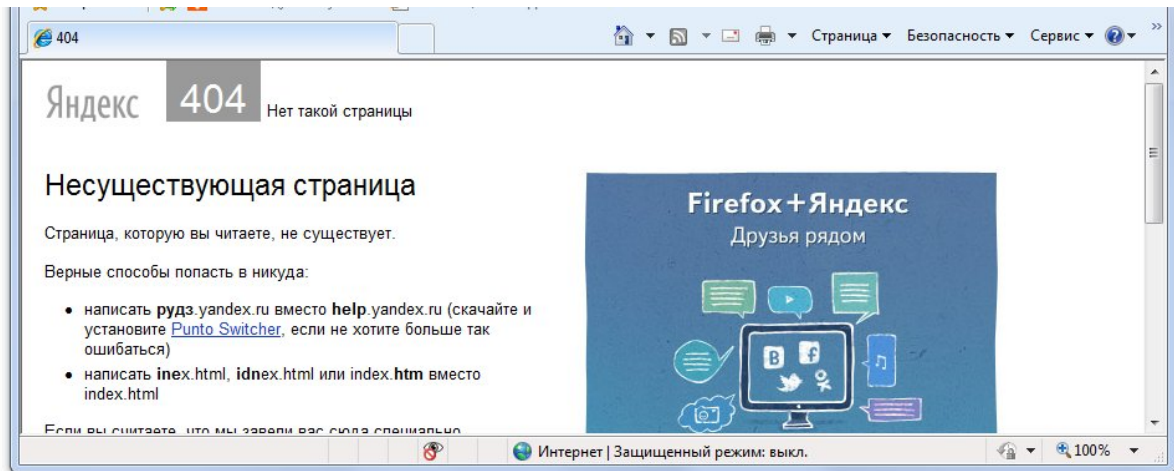
4. Так же само найдем код таких ссылок на другой странице, где они активные. Например на предыдущей и сравним код:



Как видим в качестве признака можем взять текст «**следующая**». Такого текста нет на последней странице, но в таком случае теряется универсальность признака. Ведь если кто-то ищет на другом языке, например на украинском, у него будет «**наступна**». Потому лучше не брать в качестве признака тот текст, который выводится пользователю. Нужно найти такой текст, который будет в тексте поисковика всегда, для всех языков. Можно взять идентификатор ссылки на следующую страницу: «****».

5. Теперь необходимо проверить, не встречается ли текст на последней странице. Заходим в код последней страницы и ищем. Не находим.

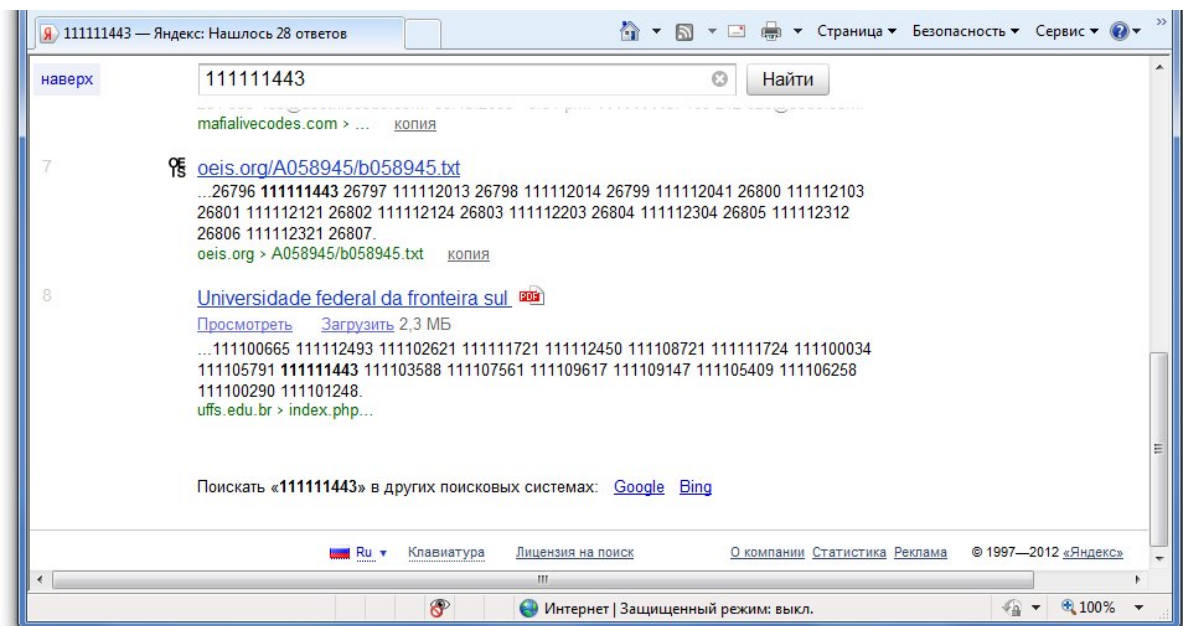
Кстати данный текст не будет присутствовать не только на последней странице, но и на всех последующих, для параметра $p=101, 102$ и т. д. Для таких запросов страница Яндекса будет выглядеть так:



Можно подбирать признак последней страницы, ориентируясь и на эту страницу.

6. Так же нужно проверить встречается ли этот текст на странице с небольшим количеством результатов (где нет ссылки на следующую страницу).

Вводим в поисковый запрос фразу, по которой вероятно существует не много ссылок, которые поместятся на одну страницу. Например «111111443»:



Поисковик нам выдал всего 8 результатов. Ссылок на следующие страницы не видно. Посмотрим HTML-код, не встречается ли текст «``» там. Как выяснилось – не встречается.

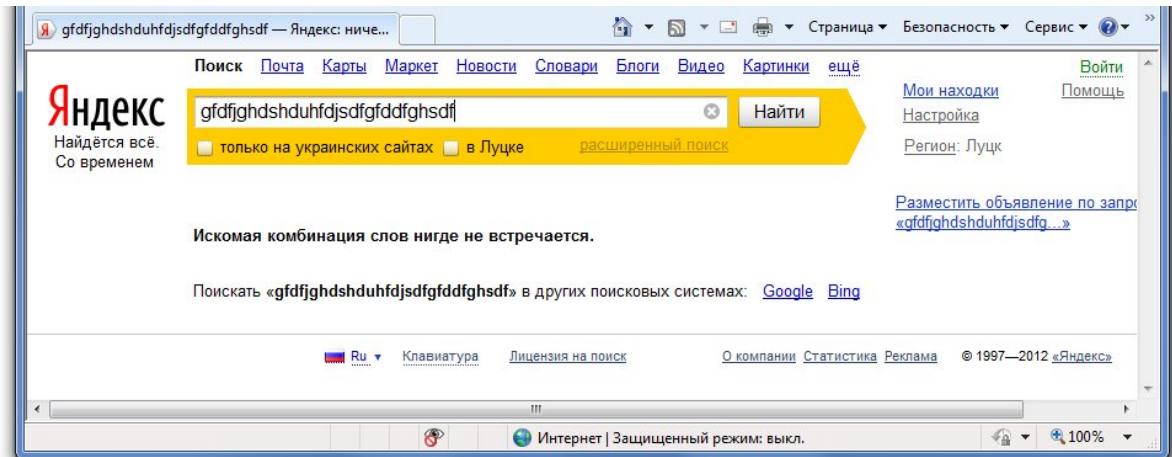
Таким образом, мы определили, что текст «``» не присутствует только на последней странице, потому мы можем его взять в качестве признака последней страницы, сняв галочку напротив поля признака.

Если бы мы нашли текст, который присутствовал только на последней странице, галочку напротив поля нужно было бы установить.

Определение признака отсутствия результатов поиска

Определение данного признака осуществляется аналогично, как и признака последней страницы.

Необходимо сформировать поисковый запрос, который не покажет ни одного поискового результата, например:



Проверяем, встречается ли наш найденный ранее текст «» в HTML-коде этой страницы. Не встречается. Потому этот текст нам подходит и в качестве признака отсутствия результатов. Вносим его туда и снимаем галочку напротив поля признака последней страницы.

Таким образом вкладка заполнения признаков профиля будет выглядеть так:

Профиль готов, можно закрыть окно и пользоваться профилем.

Для каждого профиля можно найти множество вариантов признаков, которые либо будут только на последних страницах, либо только там и не будут, но если эти условия выполняются и правильно выставлены галочки, профиль будет стабильно работать в любом варианте.