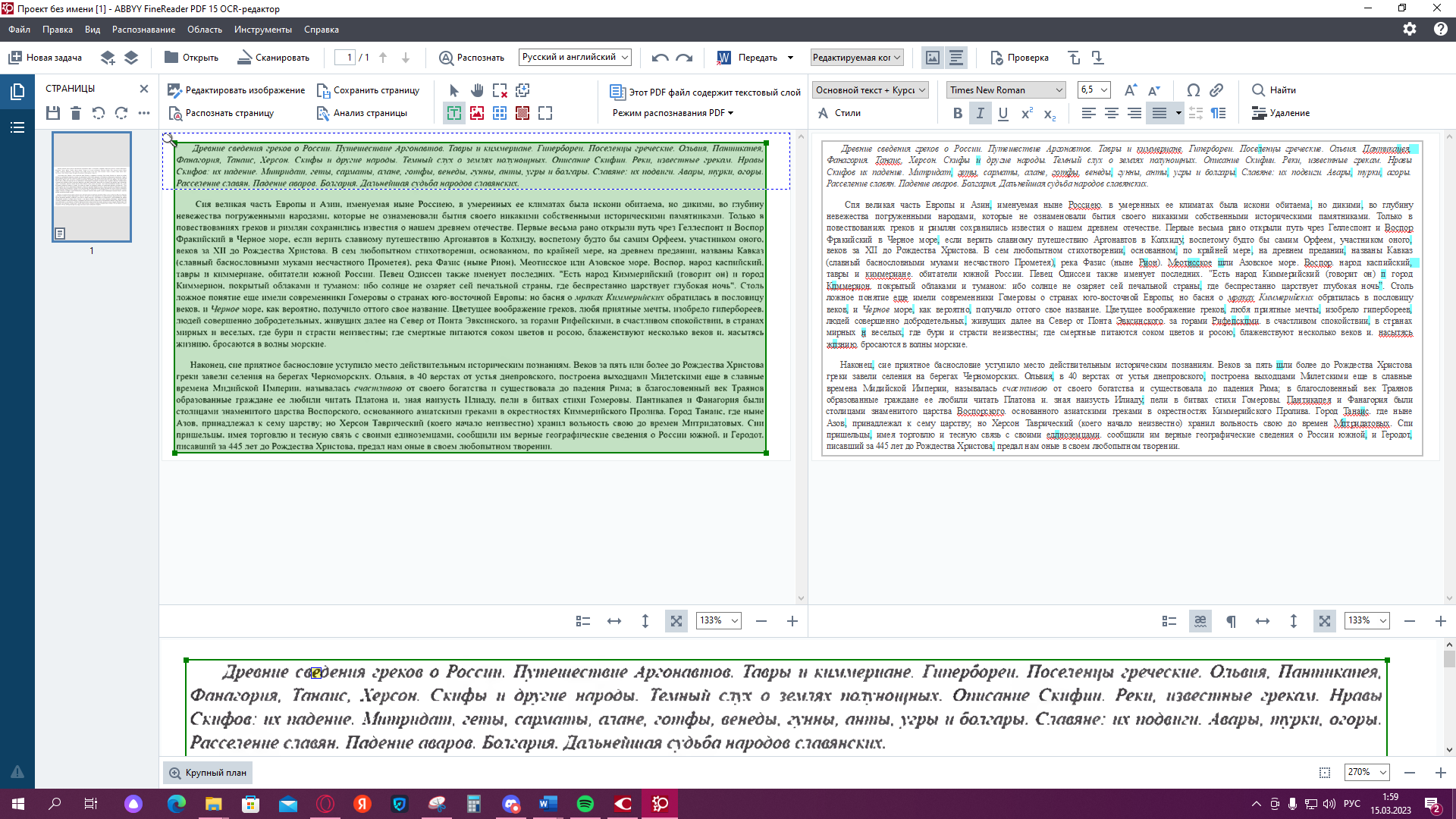
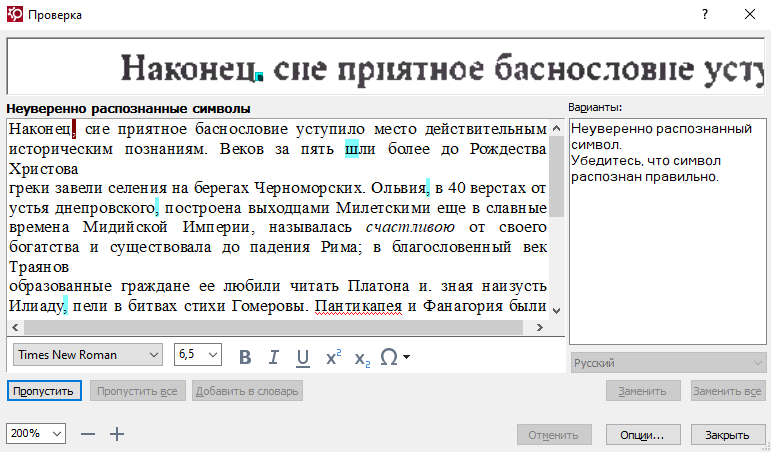
**Уровень 1**

**Задание 1.**

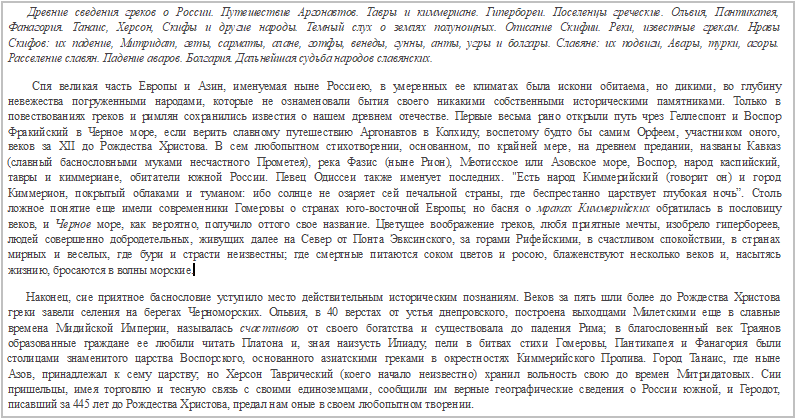
Я закинул необходимый текст в программу ABBYY Fine Reader 15, программа распознала все слова, что показано зелёным цветом слева, но не во всех была уверена, что показано голубым цветом справа.



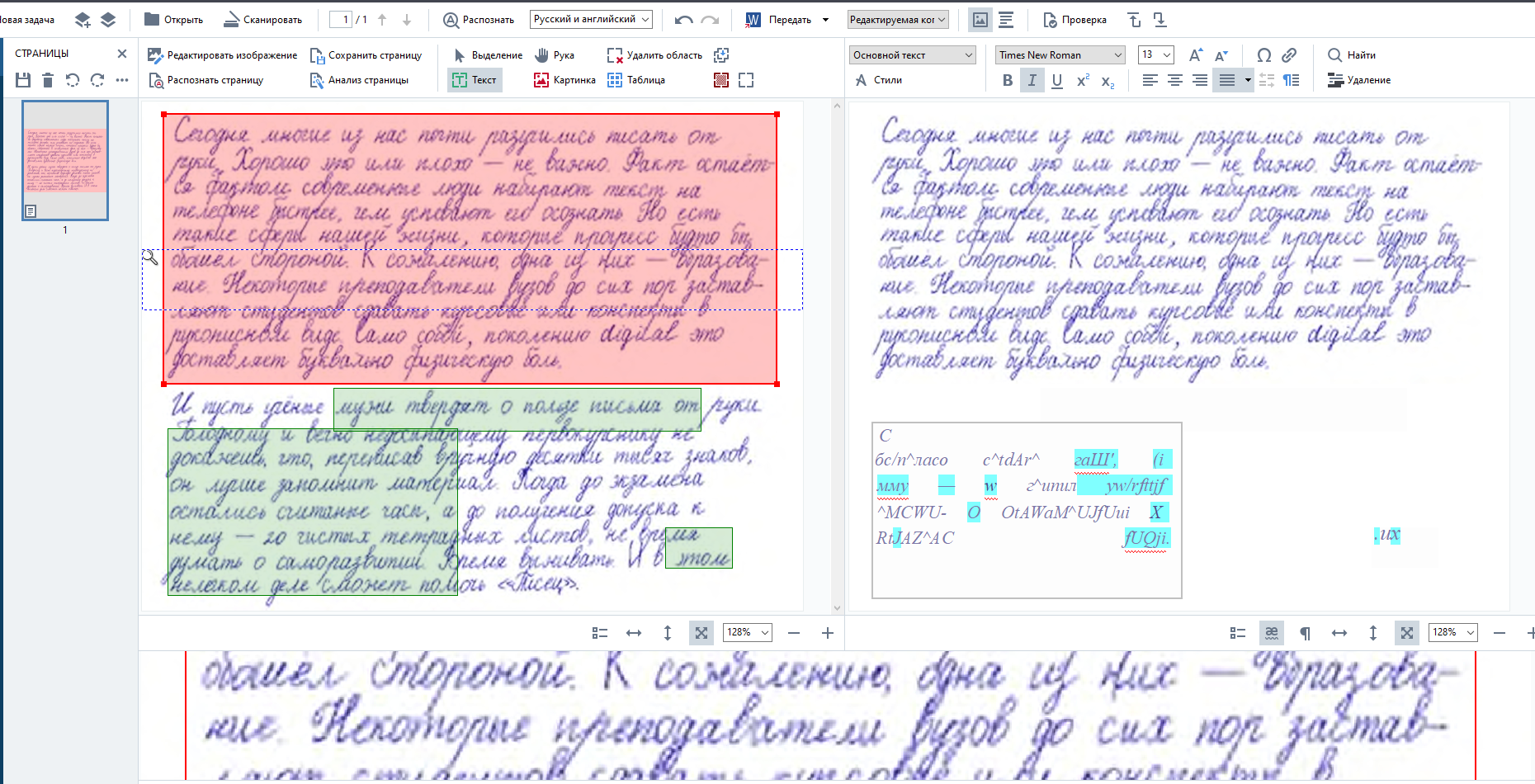
Окно проверки слов, возможно неверно распознанных программой:



Исправленный вариант:

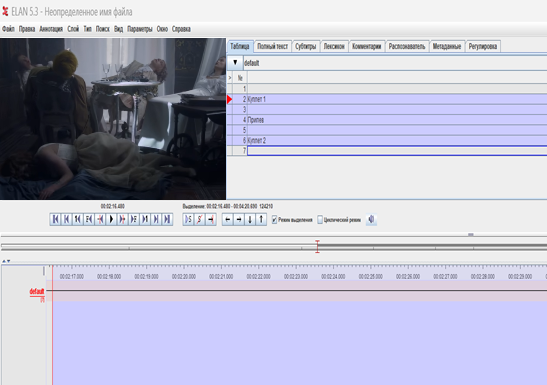


Далее я работал с рукописным текстом, как мы видим, программа вовсе не распознала текст, написанный от руки:



Исправлять текст нет смысла, ведь исправлять толком и нечего.

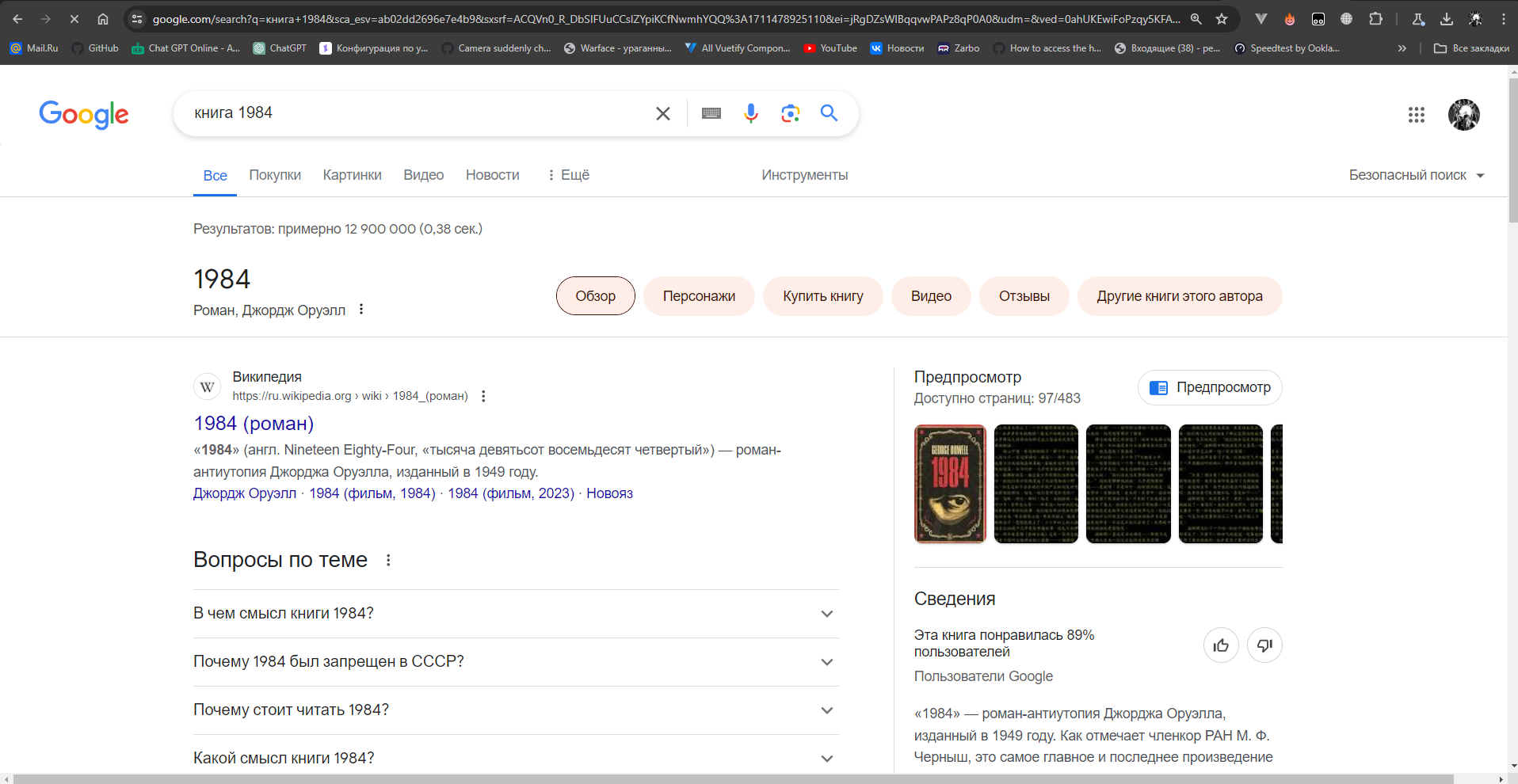
Далее я работал в программе ELAN с видео. Загрузив в него клип в формате .mp4 я сделал аннотацию песни, разделив её на несколько куплетов и припев:



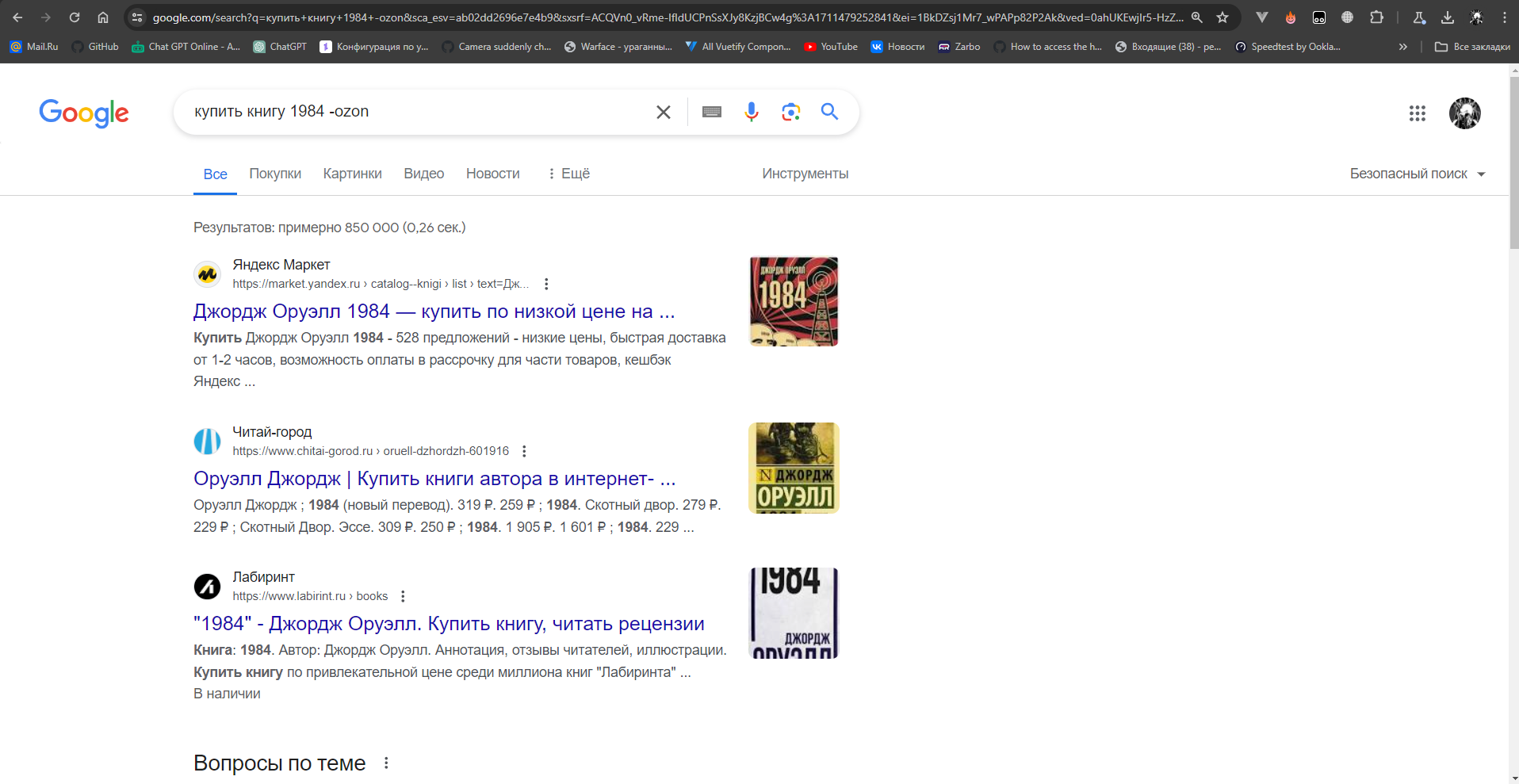
**Задание 2.**

В поисковом операторе я ввёл запросы в разных поисковых операторах:

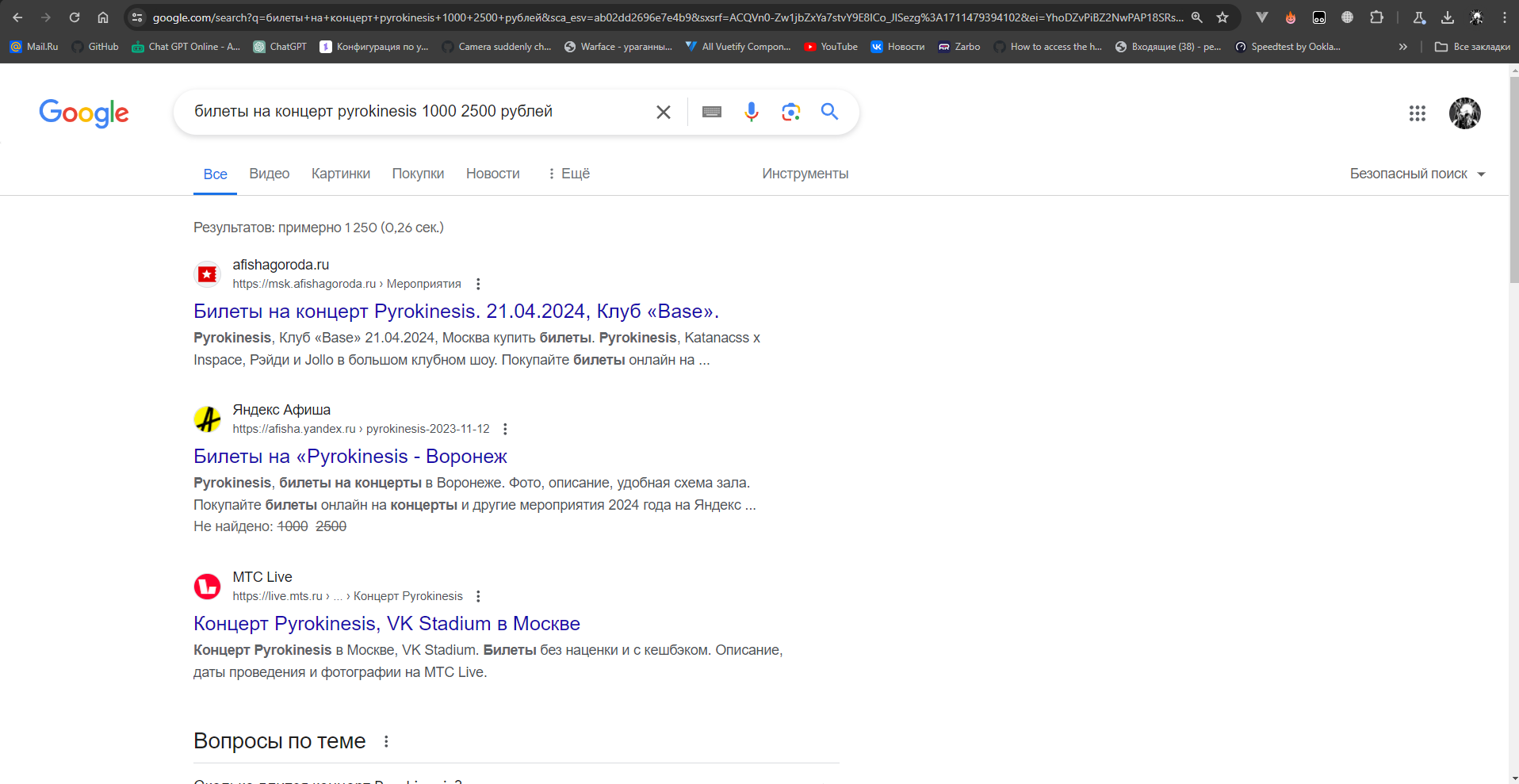
1. Точное соответствие



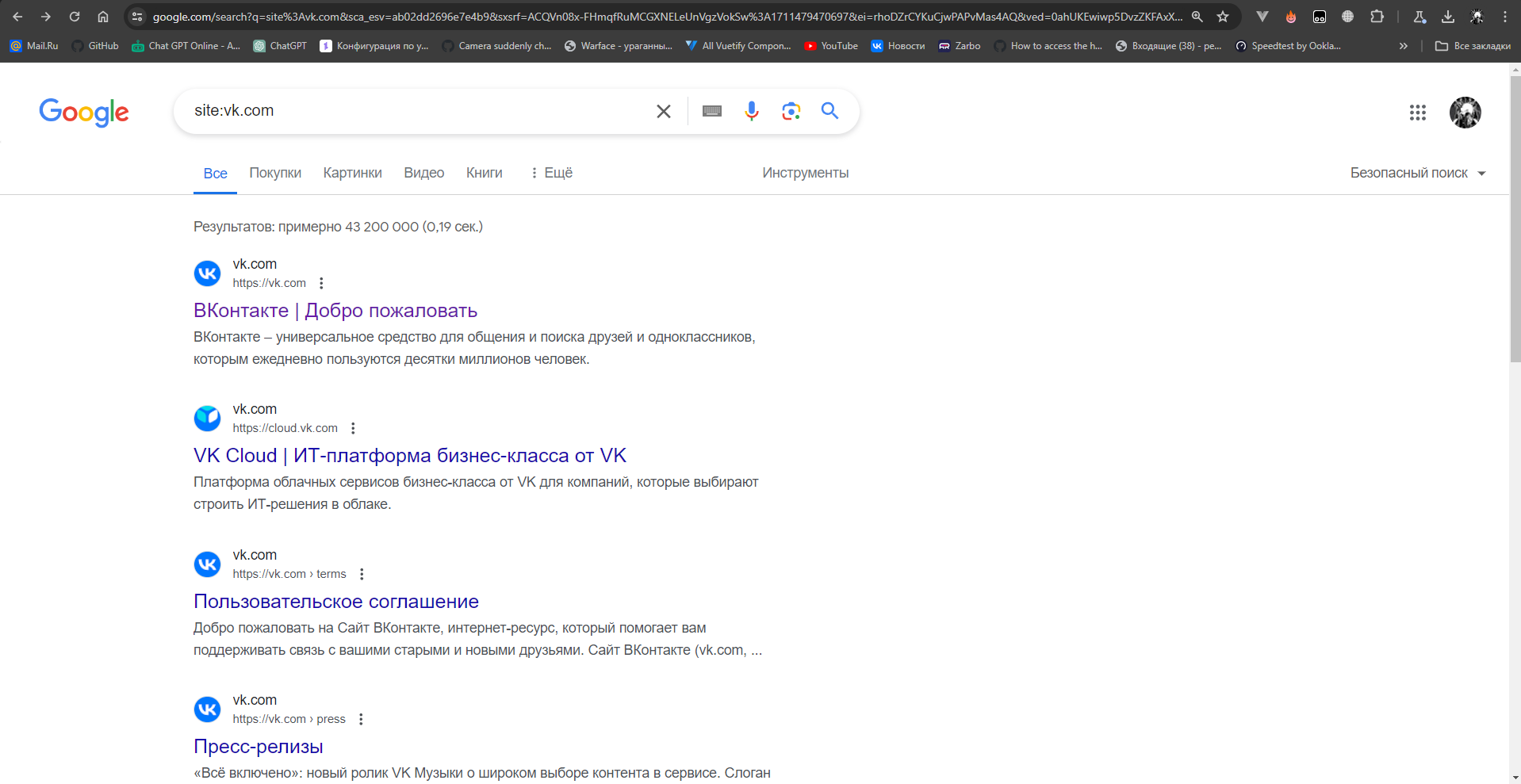
1. Исключение слов



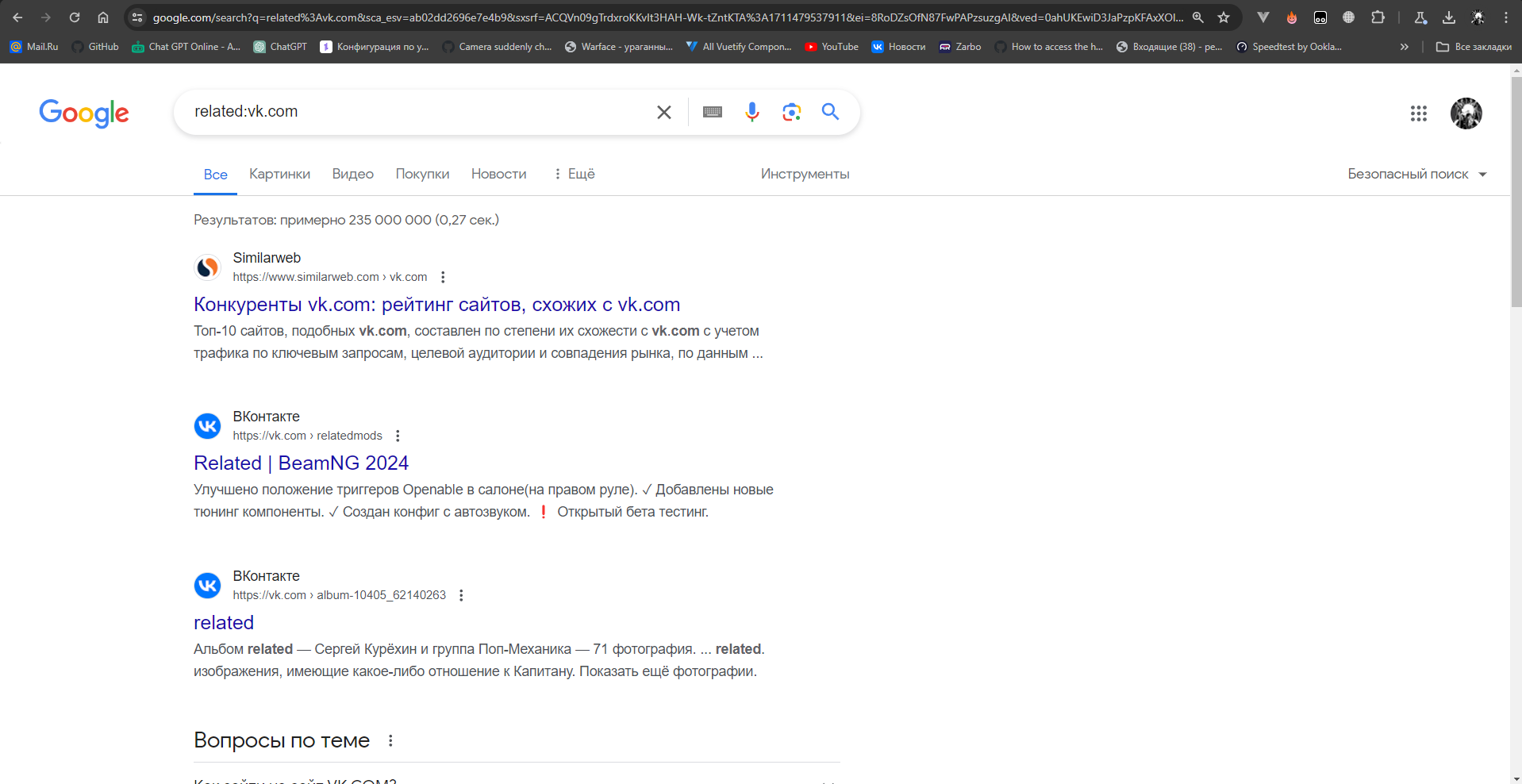
1. Задание диапазона



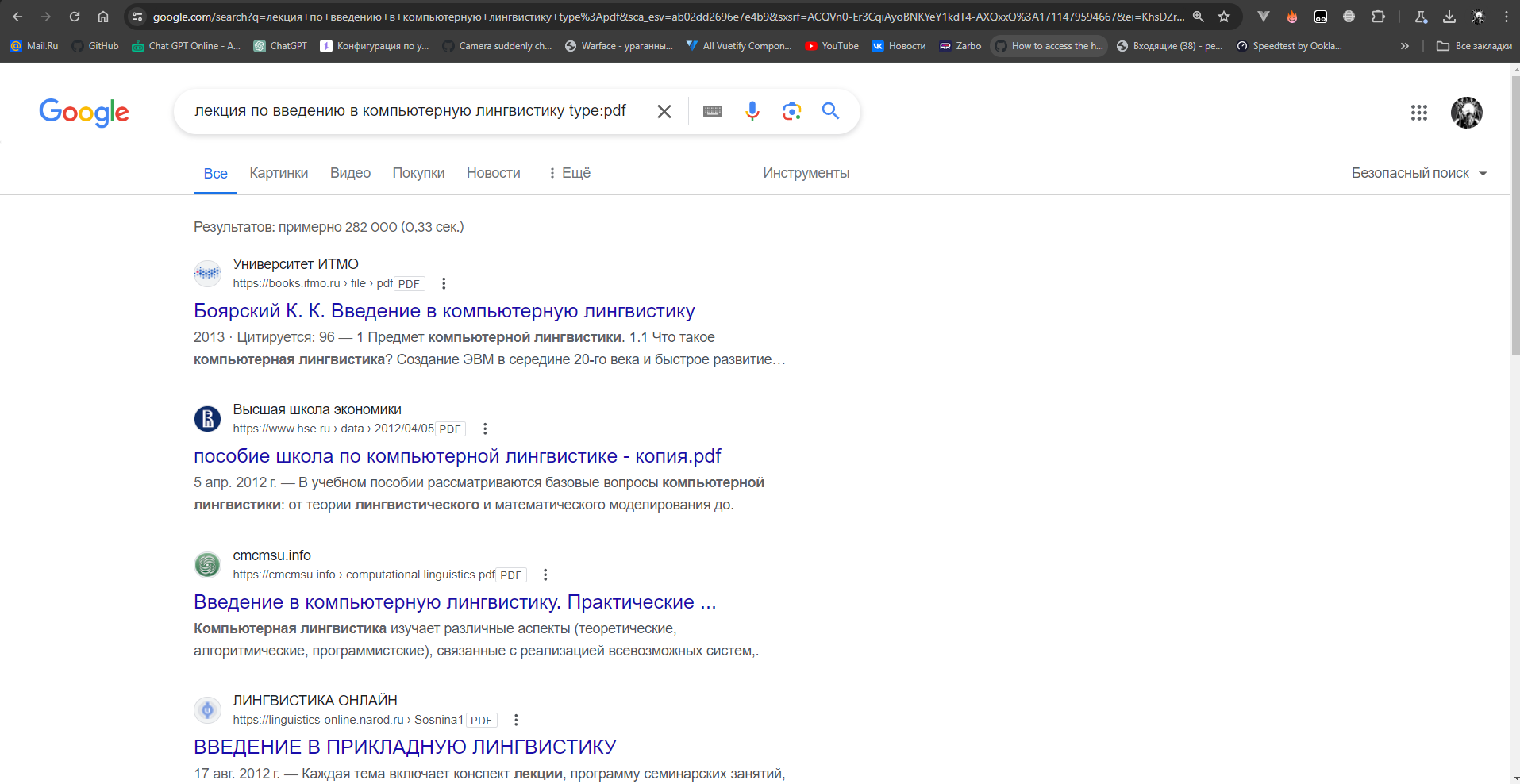
1. Поиск по конкретному сайту



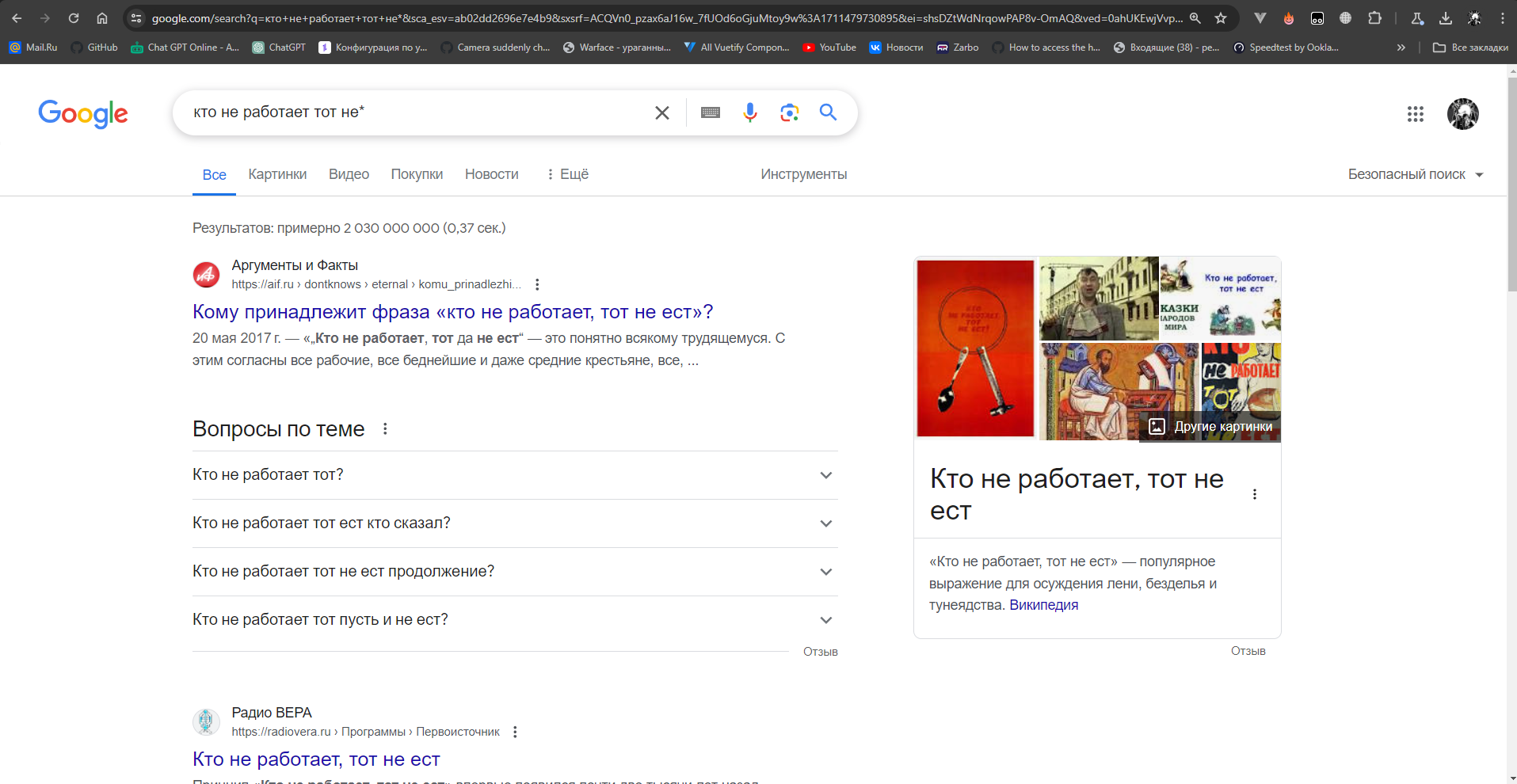
1. Поиск сайтов, похожих на известный



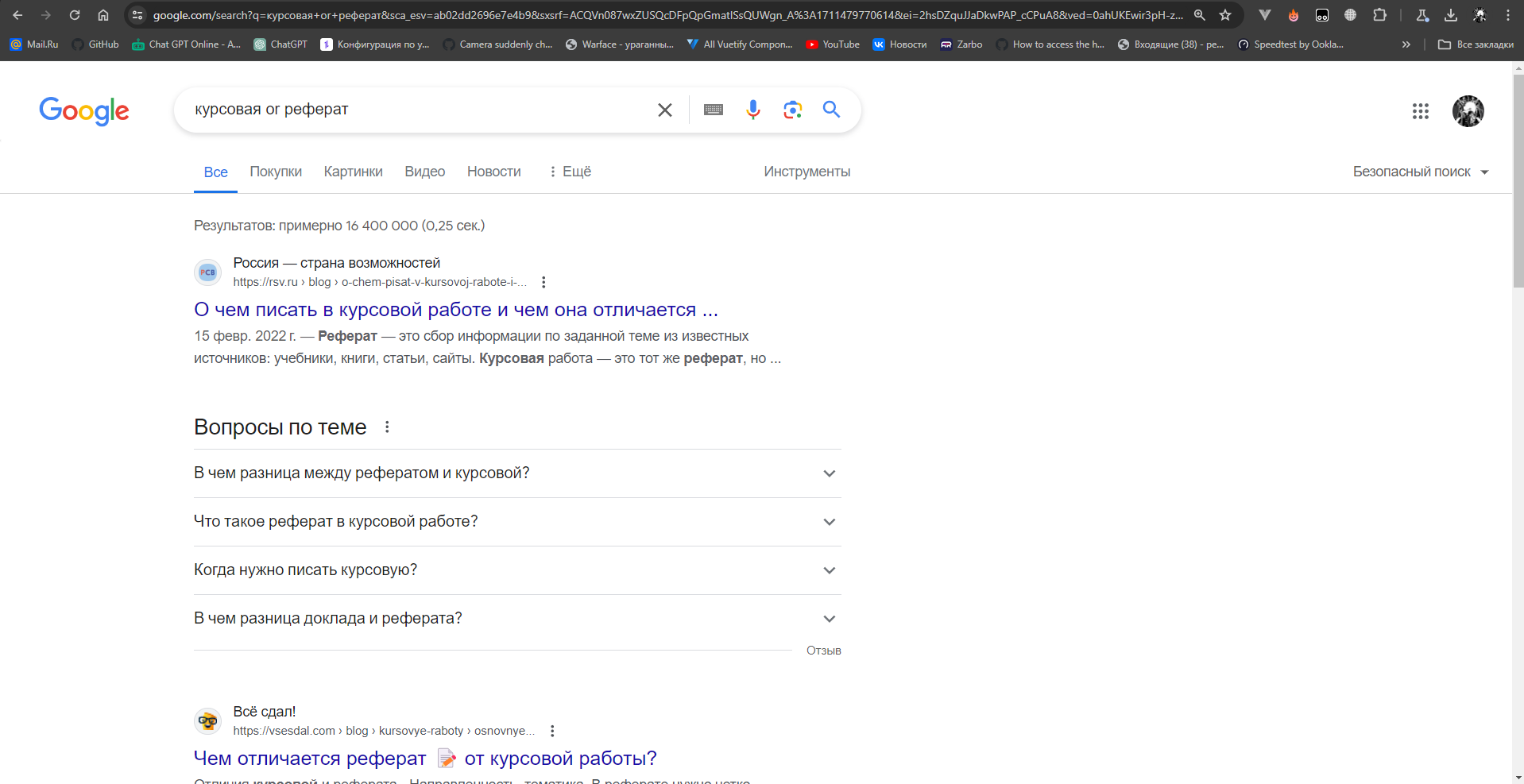
1. Поиск по формату файла



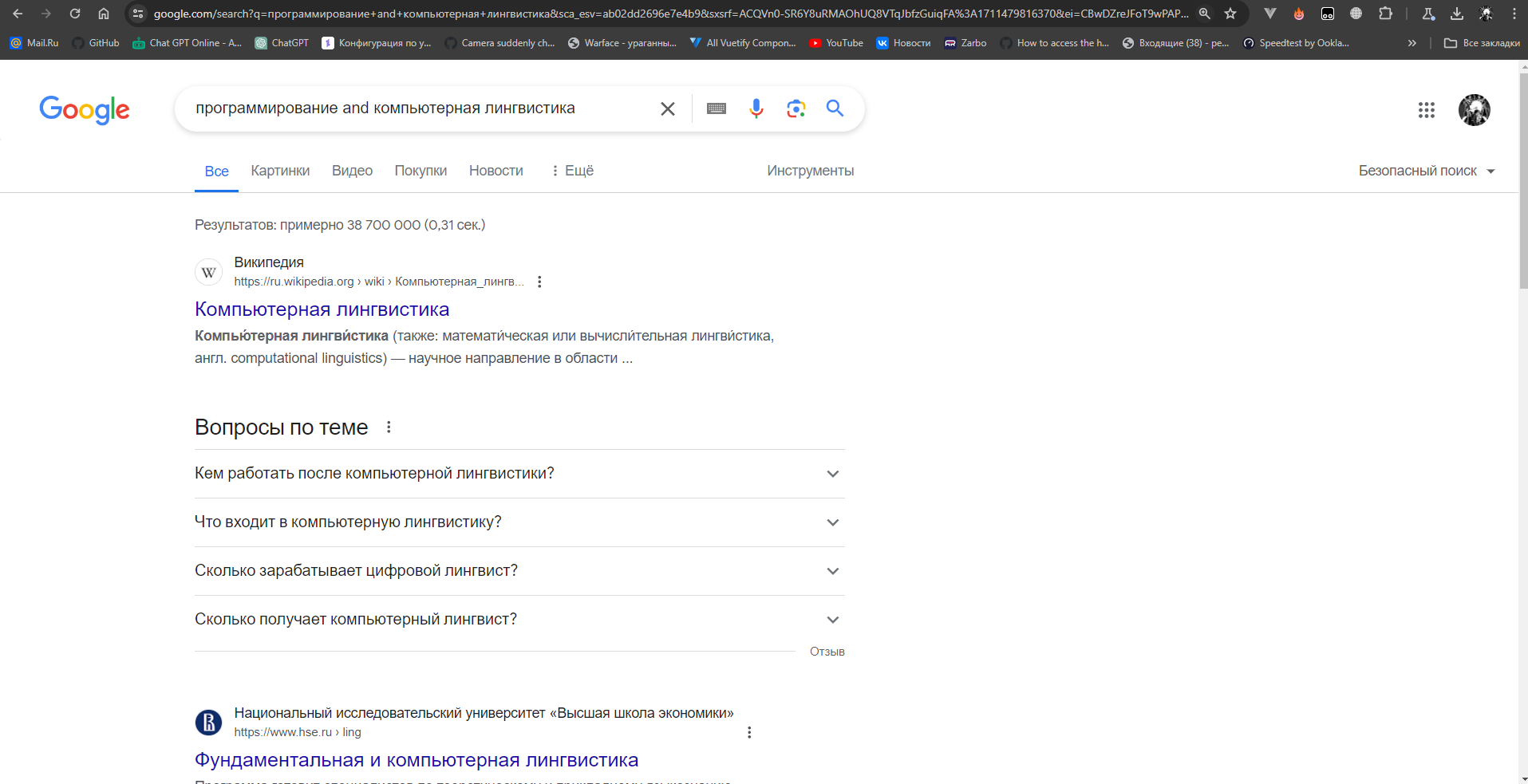
1. Восстановление пропущенного слова



1. Логическое “ИЛИ”

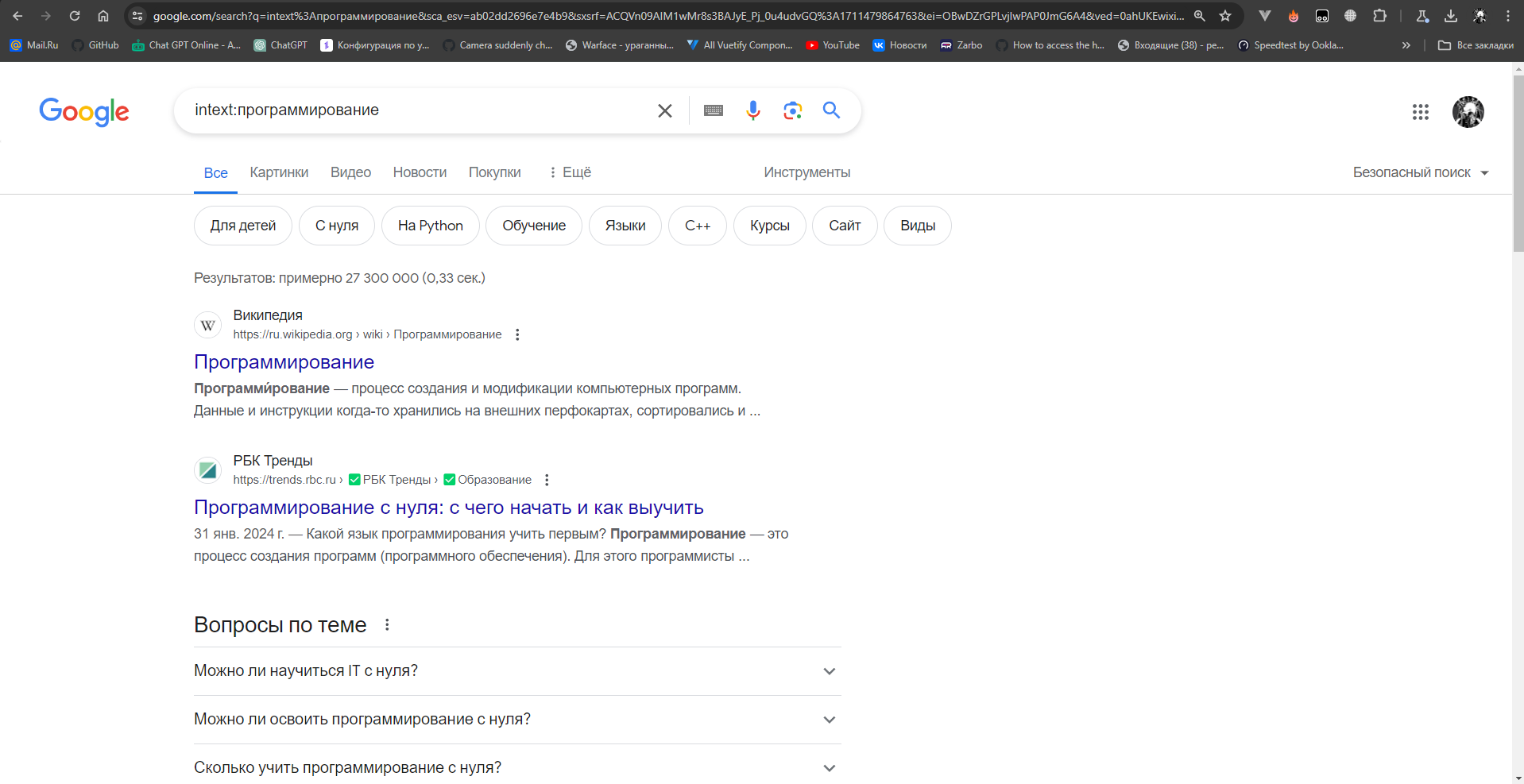


1. Логическое “И”

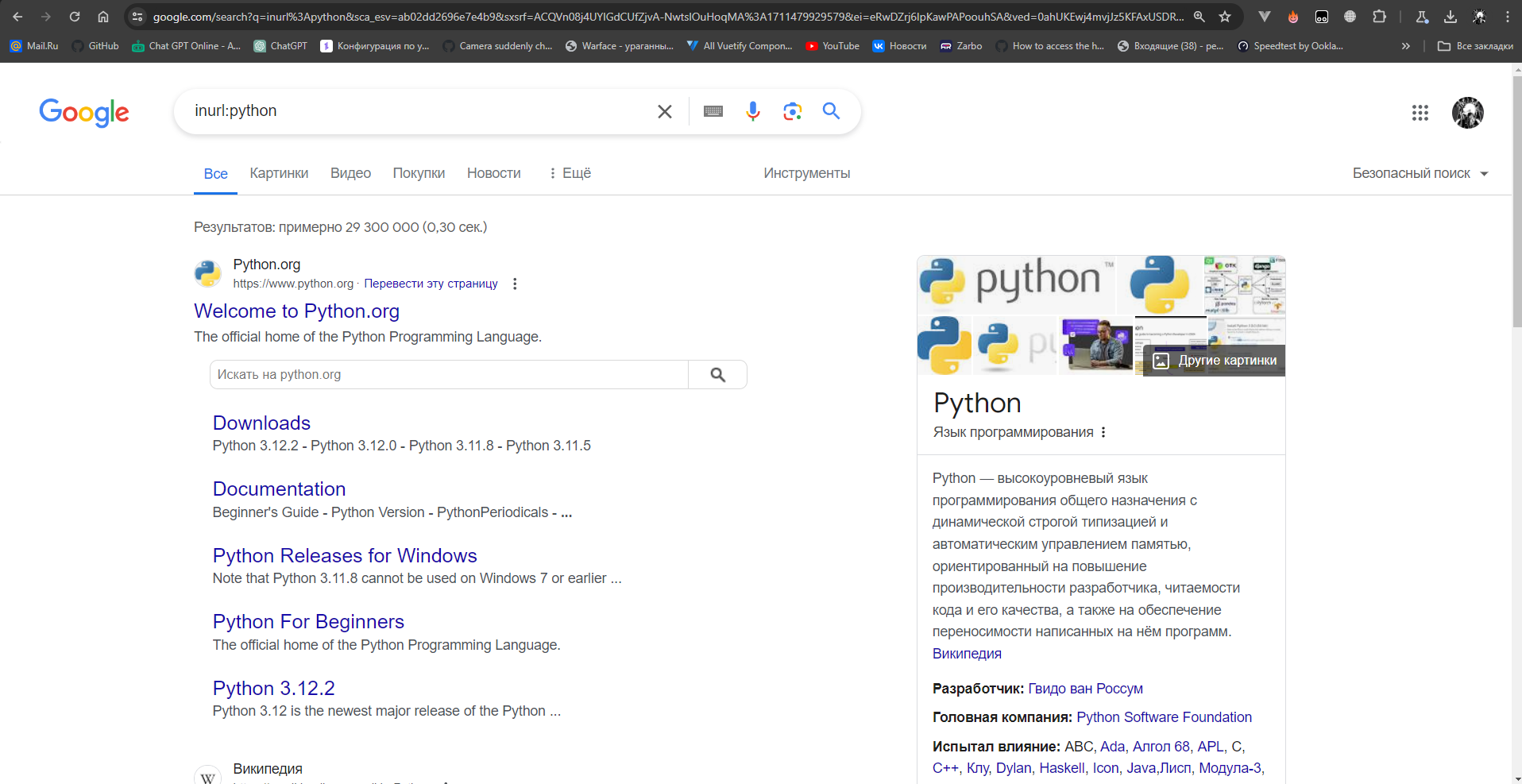


В общем и целом, результатами я остался доволен, лишь 9 поисковый оператор подвёл, оказалось не так много нужных и полезных ссылок.

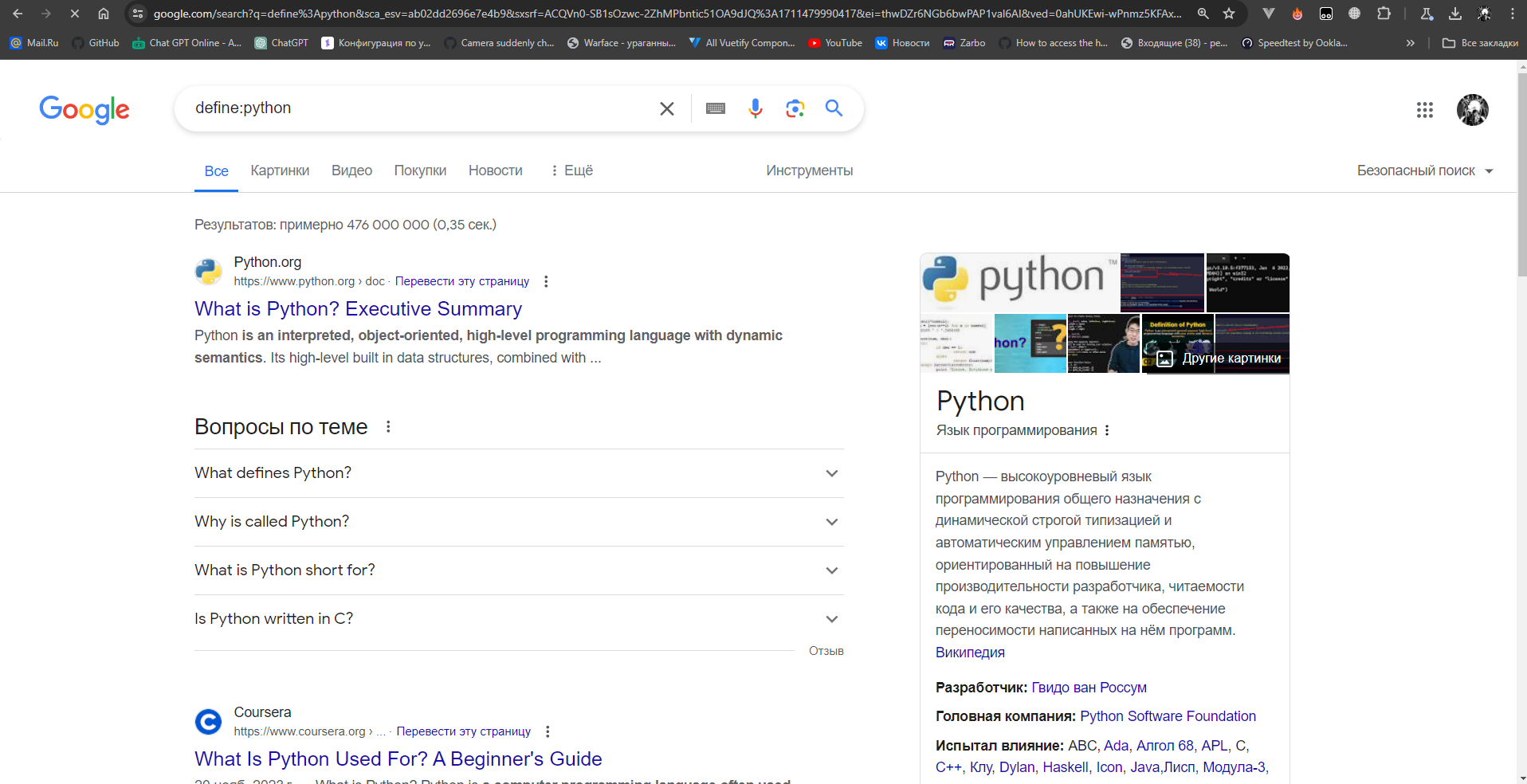
1. intext: - ищет ключевое слово в тексте страницы, игнорируя заголовки, ссылки и т. д.



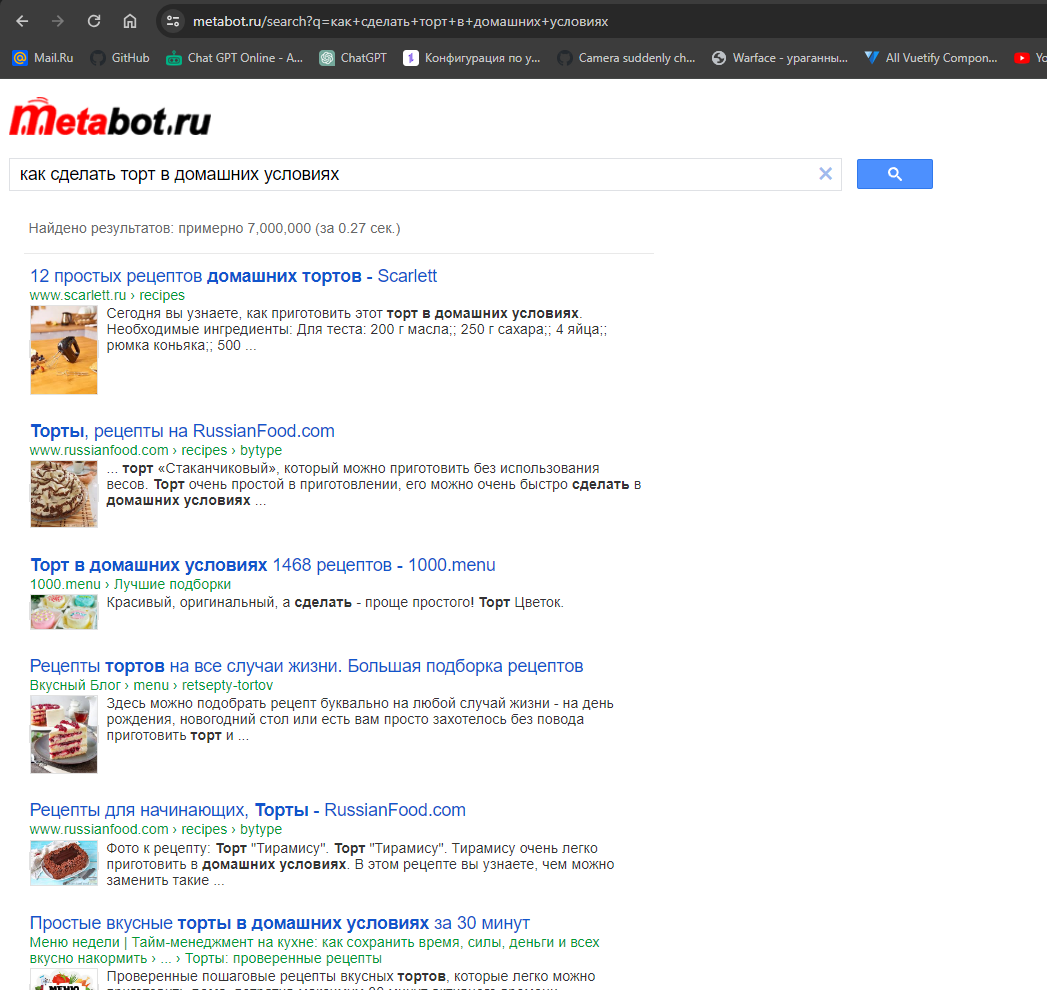
1. inurl: - ищет ключевое слово в адресе страницы.

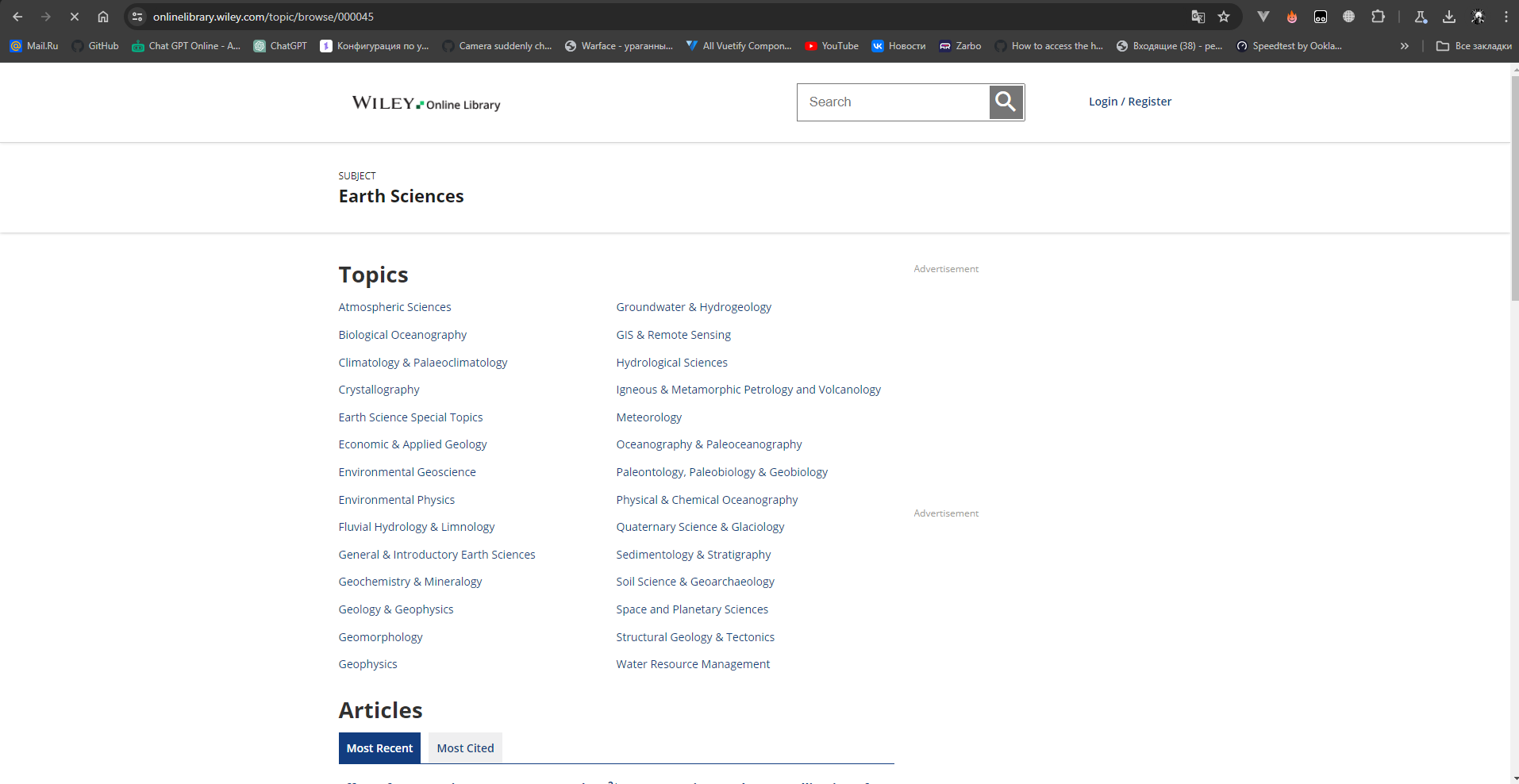


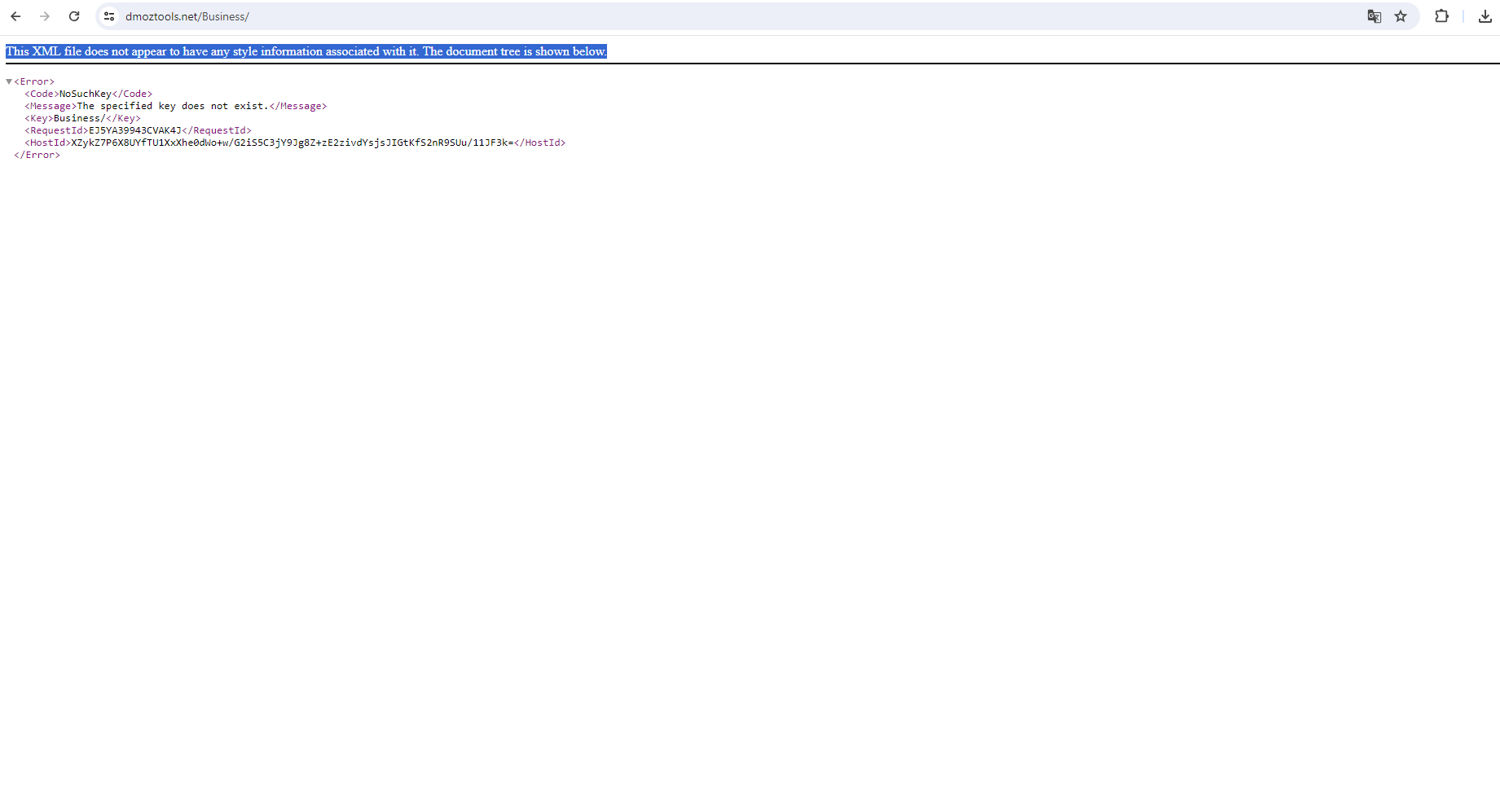
1. define: - показывает определение слова.



**Задание 3.**





В случаи с Dmoz его не получилось запустить ругается на XML файл  


Метапоисковые системы могут обеспечить более широкий диапазон результатов поиска, включая результаты из разных источников, таких как поисковые системы, каталоги, базы данных, социальные сети и другие источники. Они также могут обеспечивать более объективные результаты поиска, поскольку не являются привязанными к какой-либо конкретной поисковой системе или источнику информации.

Каталог, в моём случаем протестированный мной Wiley, является хорошим многоязычным справочником, содержит в себе множество ссылок из всемирной паутины. Каталог был организован в древовидную структуру, что делало поиск более эффективным, удобным и практичным.

Помимо рекомендаций для студентов по поиску информации в интернете, представленных в файле, могу добавить ещё от себя:

* Желательно использование нескольких поисковых систем и баз данных для получения более полной картины на тему, которую вы ищете. Также рекомендуется использовать проверенные источники информации, такие как научные журналы, учебные пособия и электронные библиотеки.
* Необходимо критически анализировать информацию, найденную в Интернете, оценивать источник, проверять его авторитетность и достоверность.
* Информация нередко спустя длительное время после публикации может терять в актуальности, поэтому необходимо проверять дату публикации и/или обновления той или иной информации.
* Использование закладок в браузере также может помочь в сохранении полезных сайтов и ссылок, часто используемых пользователем.

**Уровень 2**

***OCR (Optical Character Recognition)*** – это технология, которая позволяет компьютеру распознавать текст на изображениях, сканах и других документах. OCR-программы обладают следующими основными возможностями:

* Распознавание текста. OCR-программы могут определять символы и слова на изображениях и документах, переводя их в текстовый формат, который можно редактировать и использовать в других приложениях.
* Обработка документов. OCR-программы могут обрабатывать большие объемы документов, извлекая текст и метаданные, такие как заголовки, номера страниц, даты и т.д.
* Поддержка различных форматов. OCR-программы могут работать с различными форматами документов, включая PDF, TIFF, JPEG, BMP и другие.
* Работа на нескольких языках. OCR-программы могут распознавать текст на различных языках, включая редкие и экзотические.
* Исправление ошибок. OCR-программы могут исправлять ошибки распознавания символов, используя алгоритмы и правила проверки правописания.
* Автоматизация. OCR-программы могут использоваться для автоматического распознавания текста в документах, например, для сортировки и архивирования документов в электронном виде.
* Интеграция с другими приложениями. OCR-программы могут интегрироваться с другими приложениями, например, с учетными системами, CRM-системами или системами управления документами, чтобы облегчить их использование и повысить эффективность работы.

*Перспективы развития OCR-программ:*

* Улучшение точности распознавания. Современные OCR-программы уже достаточно точно распознают текст, но возможно, что в будущем точность распознавания будет еще выше благодаря использованию новых алгоритмов и моделей машинного обучения.
* Распознавание изображений. Возможно, что OCR-технологии будут развиваться в сторону распознавания не только текста, но и изображений.
* Распознавание голоса: Одним из перспективных направлений развития OCR-технологий является распознавание голоса, что может привести к созданию новых инструментов для анализа и обработки аудио-данных.
* Использование в различных отраслях. OCR-технологии могут найти применение во многих отраслях, включая медицину, финансы, право, образование и другие, для автоматизации процессов и ускорения работы.
* Распознавание рукописного текста. Распознавание рукописного текста - это сложная задача, но возможно, что в будущем OCR-технологии будут развиваться в этом направлении, что может упростить работу с документами, написанными от руки.

***Интеллектуальное распознавание*** – это более продвинутый подход к распознаванию текста (OCR), который включает в себя не только механическое сканирование текста с использованием оптического сканера, но и анализ и интерпретацию содержимого текста с помощью методов машинного обучения и искусственного интеллекта. В отличие от обычных OCR-программ, интеллектуальное распознавание может анализировать контекст, предназначение и содержание текста, а также определять определенные ключевые слова и понятия. Интеллектуальное распознавание может использовать различные методы машинного обучения, такие как нейронные сети, для распознавания и классификации текстовых данных, а также анализа контекста и содержания. Это позволяет создавать более точные и полезные результаты, которые могут быть использованы для автоматизации процессов и повышения эффективности работы.

Одной из систем автоматического распознавания текста является Tesseract OCR – это свободно распространяемый программный продукт с открытым исходным кодом, разработанный компанией Google. Он предназначен для распознавания текста на изображениях и в PDF-документах, используя методы машинного обучения и статистического анализа.

Особенности системы Tesseract OCR включают в себя:

* Большой объем поддерживаемых языков
* Автоматическое обнаружение языка
* Высокая точность распознавания
* Возможность использования на разных платформах
* Открытый исходный код

| Задание | Ответ |
| --- | --- |
| 1. Сколько раз в тексте встречается слово child (в разных формах)? | 22 |
| 2. Сколько раз в тексте встречается слово child именно в этой форме? | 10 |
| 3. Приведите один из контекстов использования в тексте слова beautiful | В песне черепахи |
| 4. В какой орфографии (британской или американской) представлен текст? | Ответ: британская  Проверочное слово: centre |

| Задание | Формула поиска | Ответ |
| --- | --- | --- |
| 1. Найдите в тексте первые пять слов, состоящих из пяти букв | <?????> | Alice, tired, twice, could, would |
| 2. Сколько в тексте шестибуквенных слов, начинающихся на букву s и заканчивающиеся на букву r? | <s[a-z]{4}r> | 16 |
| 3. Найдите в тексте первые пять трёхбуквенных слов, начинающиеся на гласную букву | <[aeiouy]??> | and, use, own, out, all |
| 4. Сколько в тексте слов, состоящих из двенадцати букв? По каким формальным признакам их можно сгруппировать? Приведите пример из каждой группы слов | <[a-z]{12}> | Ответ: 31  Группы: по части речи: question (noun) |
| 5. Сколько в тексте слов с суффиксом -tion? Приведите пример использования такого слова в контексте | tion> | Ответ: 59  Пример: The first question of course was… |
| 6. Есть ли в тексте слова, включающие четыре согласные буквы подряд? | [bcdfghjklmnpqrstvwxz]{4} | Да, 23 различных слова |
| 7. Сколько раз в тексте встречаются пассивные конструкции единственного числа прошедшего времени? | <was@[a-z]\*ed> | 18 |

1. Соответствие документа содержанию информационного запроса называют смысловой релевантностью, а соответствие поискового образа этого документа формализованному поисковому предписанию, выражающему данный информационный запрос, — формальной релевантностью. В целом, в то время как формальная релевантность – это простой и понятный подход к поиску, смысловая релевантность может обеспечить более точные и полезные результаты, принимая во внимание контекст и значение поискового запроса.
2. Степень соответствия информационного запроса реальной информационной потребности называется ***пертинентностью***. Способы снижения пертинентности: изменение запроса, использование альтернативных поисковых систем, уточнение запроса, использование дополнительных инструментов.
3. В ***документальных*** ИПС хранятся тексты документов или их описания (рефераты, библиографические карточки и пр.). До последнего времени обычной формой представления данных в документальных ИПС был реферат (или другое краткое описание документа) и его библиографические данные. В этом массиве, который называется первым документальным контуром, и проводился основной поиск. ***Фактографические*** ИПС имеют дело с описанием конкретных фактов, причем не обязательно в текстовой форме. Это могут быть таблицы, формулы и пр. Существуют и смешанные ИПС, включающие как документы, так и фактографическую информацию. В настоящее время фактографические ИПС строятся на основе технологий баз данных (БД).
4. Ручное индексирование подразумевает, что человек (индексатор) просматривает документы и ручным образом определяет ключевые слова, связанную с документом. После этого, индексатор добавляет эту информацию в базу данных ИПС. ***Ручное индексирование*** является более точным, так как люди могут учесть различные контексты и значения слов, но оно более трудозатратное и дорогостоящее. ***Автоматическое индексирование*** основано на использовании алгоритмов и компьютерных программ, которые анализируют содержимое документов и автоматически определяют ключевые слова. Автоматическое индексирование позволяет быстро индексировать большое количество документов, но может быть менее точным, особенно при работе с сложными и многозначными терминами. Некоторые примеры систем ручного индексирования:

* MEDLINE - база данных медицинских статей, вручную индексируемая экспертами-библиографами.
* INSPEC - база данных по инженерии и техническим наукам, которая также вручную индексируется экспертами.

Некоторые примеры систем автоматического индексирования:

* Google - автоматически индексирует миллионы документов в Интернете, используя свои алгоритмы.
* Microsoft Academic - автоматически индексирует миллионы научных статей и других исследований с помощью своих алгоритмов.

В современных ИПС часто используется комбинированный подход, где ручное и автоматическое индексирование используются вместе для достижения наилучшей точности и скорости работы.

1. ***Общий каталог веб-ресурсов*** – это интернет-ресурс, на котором собраны ссылки на различные веб-сайты по разным тематикам, без определенной специализации. Такой каталог помогает пользователям быстро находить нужные сайты в разных категориях, но не специализируется на какой-то одной конкретной теме. Примеры общих каталогов веб-ресурсов: AllTop, Reddit.

***Специализированный каталог веб-ресурсов*** – это каталог, который специализируется на конкретной тематике или отрасли. Такие каталоги помогают пользователям быстро найти ресурсы по определенной теме или в конкретной отрасли. Примеры специализированных каталогов веб-ресурсов: GitHub, Airbnb, Tripadvisor.

1. ***Фасетная классификация*** – это метод классификации информации, основанный на описании документов несколькими независимыми характеристиками, называемыми фасетами. Фасеты позволяют пользователям уточнять поиск, добавляя дополнительные фильтры и условия. Пример фасетов:

Фасет "Бренд": Apple

Фасет "Модель": iPhone 13

Фасет "Цвет": графитовый

Фасет "Ёмкость памяти": 256 ГБ

Фасет "Цена": 71990 рублей

**Google:**

**Запрос:** где и когда появился термин «лингвистика»

**Документ, отвечающий результатам запроса (url):** http://slovar.lib.ru/dictionary/lingvistika.htm#:~:text=Lingua%20-%20язык)%20-%20наука%2C,"Словаря%20французского%20языка"%20Ш.

**Номер этого документа в списке результатов:** 1

**Инф. шум (количество нерелевантных ссылок):** около 4 ссылок отвечало конкретно на мой запрос, однако было много смежных ссылок, на которых был представлен либо более развёрнутый ответ, либо просто описывалось определение лингвистики.

**Полнота (в Рунете всего 4 источника):** буквально по первой выданной ссылке был получен исчерпывающий полный ответ без какой-либо воды.

**Рамблер:**

**Запрос:** где и когда появился термин «лингвистика»

**Документ, отвечающий результатам запроса (url):** https://znachenie-slova.ru/лингвистика

**Номер этого документа в списке результатов:** 3

**Инф. шум (количество нерелевантных ссылок):** только две ссылки с ближайших страниц отвечали конкретно на поставленный вопрос, остальные в основном описывали сам предмет лингвистики.

**Полнота (в Рунете всего 4 источника):** даже одного источника хватило на то, чтобы понять, что информация дана достаточно полная и исчерпывающая, хоть и есть немного лишнего, чего не было в моём запросе.

Однозначно я бы предпочёл ИПС Google, ведь там почти все первые ссылки отвечали моему запросу, было намного меньше информационного шума, нежели в Рамблере. В удобстве выделить никого не могу, задачи и функционал у них одинаковый. Визуально не нравится не один, глаз не цепляет, хотя оформление, запоминаемость и красота – далеко не первая задача поисковых систем.