**Уровень 1**

**файл “регулярные выражения”**

1. ([0-1][0-9]|2[0-3]):[0-5][0-9]
2. В строке Java она обнаружит Java, а в строке JavaScript - JavaS, т.к. регулярное выражение ищет Java и любой символ, кроме тех что стоят в списке, а там представлено s в нижнем регистре.
3. #(\d|[A-F])(\d|[A-F])(\d|[A-F])(\d|[A-F])(\d|[A-F])(\d|[A-F])
4. (-?\d+\.?\d\*)\s\*([+]|[-]|[\*]|[/])\s\*(-?\d+\.?\d\*)

**файл “Форматы данных и кодировки. Макросы в Sublime**

Шифровка текста через сайт:

Исходный текст:

**Язык разметки** — набор символов или последовательностей, вставляемых в текст для передачи информации о его выводе или строении. Текстовый документ, написанный с использованием языка разметки, содержит не только сам текст (как последовательность слов и знаков препинания), но и дополнительную информацию о различных его участках — например, указание на заголовки, выделения, списки и т. д. В более сложных случаях язык разметки позволяет вставлять в документ интерактивные элементы и содержание других документов. К языкам разметки относятся, например, XML и HTML.

Закодированное сообщение:

ذ¯ذ·ر‹ذ؛ ر€ذ°ذ·ذ¼ذµر‚ذ؛ذ¸ â€” ذ½ذ°ذ±ذ¾ر€ رپذ¸ذ¼ذ²ذ¾ذ»ذ¾ذ² ذ¸ذ»ذ¸ ذ؟ذ¾رپذ»ذµذ´ذ¾ذ²ذ°ر‚ذµذ»رŒذ½ذ¾رپر‚ذµذ¹, ذ²رپر‚ذ°ذ²ذ»رڈذµذ¼ر‹ر… ذ² ر‚ذµذ؛رپر‚ ذ´ذ»رڈ ذ؟ذµر€ذµذ´ذ°ر‡ذ¸ ذ¸ذ½ر„ذ¾ر€ذ¼ذ°ر†ذ¸ذ¸ ذ¾ ذµذ³ذ¾ ذ²ر‹ذ²ذ¾ذ´ذµ ذ¸ذ»ذ¸ رپر‚ر€ذ¾ذµذ½ذ¸ذ¸. ذ¢ذµذ؛رپر‚ذ¾ذ²ر‹ذ¹ ذ´ذ¾ذ؛رƒذ¼ذµذ½ر‚, ذ½ذ°ذ؟ذ¸رپذ°ذ½ذ½ر‹ذ¹ رپ ذ¸رپذ؟ذ¾ذ»رŒذ·ذ¾ذ²ذ°ذ½ذ¸ذµذ¼ رڈذ·ر‹ذ؛ذ° ر€ذ°ذ·ذ¼ذµر‚ذ؛ذ¸, رپذ¾ذ´ذµر€ذ¶ذ¸ر‚ ذ½ذµ ر‚ذ¾ذ»رŒذ؛ذ¾ رپذ°ذ¼ ر‚ذµذ؛رپر‚ (ذ؛ذ°ذ؛ ذ؟ذ¾رپذ»ذµذ´ذ¾ذ²ذ°ر‚ذµذ»رŒذ½ذ¾رپر‚رŒ رپذ»ذ¾ذ² ذ¸ ذ·ذ½ذ°ذ؛ذ¾ذ² ذ؟ر€ذµذ؟ذ¸ذ½ذ°ذ½ذ¸رڈ), ذ½ذ¾ ذ¸ ذ´ذ¾ذ؟ذ¾ذ»ذ½ذ¸ر‚ذµذ»رŒذ½رƒرژ ذ¸ذ½ر„ذ¾ر€ذ¼ذ°ر†ذ¸رژ ذ¾ ر€ذ°ذ·ذ»ذ¸ر‡ذ½ر‹ر… ذµذ³ذ¾ رƒر‡ذ°رپر‚ذ؛ذ°ر… â€” ذ½ذ°ذ؟ر€ذ¸ذ¼ذµر€, رƒذ؛ذ°ذ·ذ°ذ½ذ¸ذµ ذ½ذ° ذ·ذ°ذ³ذ¾ذ»ذ¾ذ²ذ؛ذ¸, ذ²ر‹ذ´ذµذ»ذµذ½ذ¸رڈ, رپذ؟ذ¸رپذ؛ذ¸ ذ¸ ر‚. ذ´. ذ’ ذ±ذ¾ذ»ذµذµ رپذ»ذ¾ذ¶ذ½ر‹ر… رپذ»رƒر‡ذ°رڈر… رڈذ·ر‹ذ؛ ر€ذ°ذ·ذ¼ذµر‚ذ؛ذ¸ ذ؟ذ¾ذ·ذ²ذ¾ذ»رڈذµر‚ ذ²رپر‚ذ°ذ²ذ»رڈر‚رŒ ذ² ذ´ذ¾ذ؛رƒذ¼ذµذ½ر‚ ذ¸ذ½ر‚ذµر€ذ°ذ؛ر‚ذ¸ذ²ذ½ر‹ذµ رچذ»ذµذ¼ذµذ½ر‚ر‹ ذ¸ رپذ¾ذ´ذµر€ذ¶ذ°ذ½ذ¸ذµ

ذ´ر€رƒذ³ذ¸ر… ذ´ذ¾ذ؛رƒذ¼ذµذ½ر‚ذ¾ذ². ذڑ رڈذ·ر‹ذ؛ذ°ذ¼ ر€ذ°ذ·ذ¼ذµر‚ذ؛ذ¸ ذ¾ر‚ذ½ذ¾رپرڈر‚رپرڈ, ذ½ذ°ذ؟ر€ذ¸ذ¼ذµر€, XML ذ¸ HTML.

файл пдф - рядом лежит на гите

конспект:

Информация на цифровых носителях хранится в закодированном виде и представляет собой последовательность 0 и 1. Как это устроено физически? В основе запоминающего устройства может лежать любой физический эффект, который обеспечивает приведение системы к двум устойчивым состояниям: одно из них будет обозначать 0, другое - 1. Так, в современных устройствах используются:

• наличие/отсутствие заряда в конденсаторе

• направление намагниченности

• отражение/рассеяние света от поверхности CD, DVD или Blu-ray диска

Минимальная единица хранения информации - 1 бит. С помощью него можно закодировать два значения - 0 и 1.

Упорядоченность информации обеспечивается файловой системой, которая определяет формат содержимого и способ его хранения. Иными словами, файловая система - это раздел жесткого диска, выделенный для хранения файлов на компьютере и другом электронном оборудовании: мобильных телефонах, фотоаппаратах и т.д. Рассмотрим ее устройство на примере FAT - файловой системы, впервые созданной в 1976-1977 годах Биллом Гейтсом и Марком МакДональдом.

Название FAT расшифровывается как File Allocation Table, или "таблица размещения файлов". Это значит, что "адрес" каждого файла на диске прописан в специальной таблице. Дисковое пространство разбито на кластеры, и каждый файл занимает не менее одного из них. Файловая система FAT состоит из следующих частей:

1) Загрузочная запись

а) BPB (BIOS parameter block) – содержит указание на тип файловой системы и физические характеристики носителя

б) FSINFO – сообщает о том, сколько кластеров свободно

2) FAT – та самая таблица соответствия файлов и кластеров, в которых они расположены

3) Собственно файлы

Обычно большие файлы записываются в кластеры, идущие подряд, но, когда свободного места на диске остается мало, нужного количества свободных ячеек подряд может не быть. Однако если суммарное количество свободных кластеров достаточно для записи файла, он записывается по кусочкам в разные места диска. Это называется фрагментацией. Правда, разбитые на кусочки файлы читаются гораздо медленнее, поэтому существует дефрагментация — перераспределение файлов на диске таким образом, чтобы каждый из них хранился в непрерывной последовательности кластеров.

Практически любая файловая система иерархична: файлы объединяются в каталоги (папки), которые образуют одно или несколько деревьев. Расположение файла в такой структуре описывается с помощью пути.

Кроме того, у файла есть расширение - несколько символов после точки, которые обозначают его тип (формат): txt, doc, jpg, mp3, html... Проще говоря, расширение указывает на то, какая информация (текстовая, графическая и т.д.) находится в файле и как правильно ее раскодировать.

Например, если мы попробуем вручную поменять расширение картинки с jpg на txt и затем открыть ее в текстовом редакторе, мы увидим "кракозябры", потому что байтовая последовательность, которая обозначает цвета пикселей, будет расшифрована неправильно – в данном случае как текстовые символы.

Дело в том, что для перевода последовательности нулей и единиц в текстовые символы используется специальная таблица, которая называется **кодировкой**; если текст был зашифрован с помощью одной кодировки, а расшифрован с помощью другой, то мы увидим кракозябры.

Кодировки:

**ASCII** (1963 год) — 7-битная кодировка, включающая в себя 128 символов: 33 непечатных управляющих символа (влияющих на обработку текста и пробелов) и 95 печатных символов, включая цифры, буквы латинского алфавита в строчном и прописном вариантах и ряд пунктуационных символов.

**ISO/IEC 646** (1972 год) — группы кодировок, основанных на ASCII, где редко используемые символы из ASCII заменялись на необходимые. Группа включала в себя варианты кодировки для Канады, Китая, Кубы, Германии, Дании, Финляндии, Франции, Великобритании, Греции, Венгрии, Ирландии, Японии, Южной Кореи, Мальты, Норвегии, Швеции и Югославии.

**KOI8** (1974 год, СССР) — Код Обмена Информацией. Как следует из названия, это была 8-битная кодировка, что позволяло включить в нее в два раза больше символов. KOI8 включала в себя цифры, буквы латинского и русского алфавита, а также знаки пунктуации, спецсимволы и псевдографику.

**ISO/IEC 8859** (середина 80-х). Ранние кодировки были ограничены 7 битами из-за особенностей некоторых протоколов передачи данных. Однако со временем эти ограничения свою актуальность потеряли, в то время как необходимость в дополнительных символах для языков, использующих латинский алфавит, только росла.

**Windows-125x** (начало 90-х) — группа кодировок, разработанная компанией Microsoft для ОС Windows. В Windows-1251 вошли все символы русского и близких к нему языков: украинского, белорусского, болгарского, сербского и македонского.

**UTF-8** (Unicode Transformation Format, 8-bit — «формат преобразования Юникода, 8-битный», 1993 год) — одна из общепринятых и стандартизированных кодировок текста, которая позволяет хранить символы Юникода, используя переменное количество байт (от 1 до 6)

**Unicode** – это стандарт кодирования символов, включающий в себя знаки почти всех письменных языков мира. В настоящее время стандарт является доминирующим в Интернете

Если кодировка сбита, то по характеру кракозябр можно понять, в какой кодировке был текст и в какой кодировке он отображается.

**Ресурсы для текст майнинга:**

## GATE

## KNIME

## Orange

## RapidMiner

## LPU

## MeaningCloud

## Google Cloud NLP

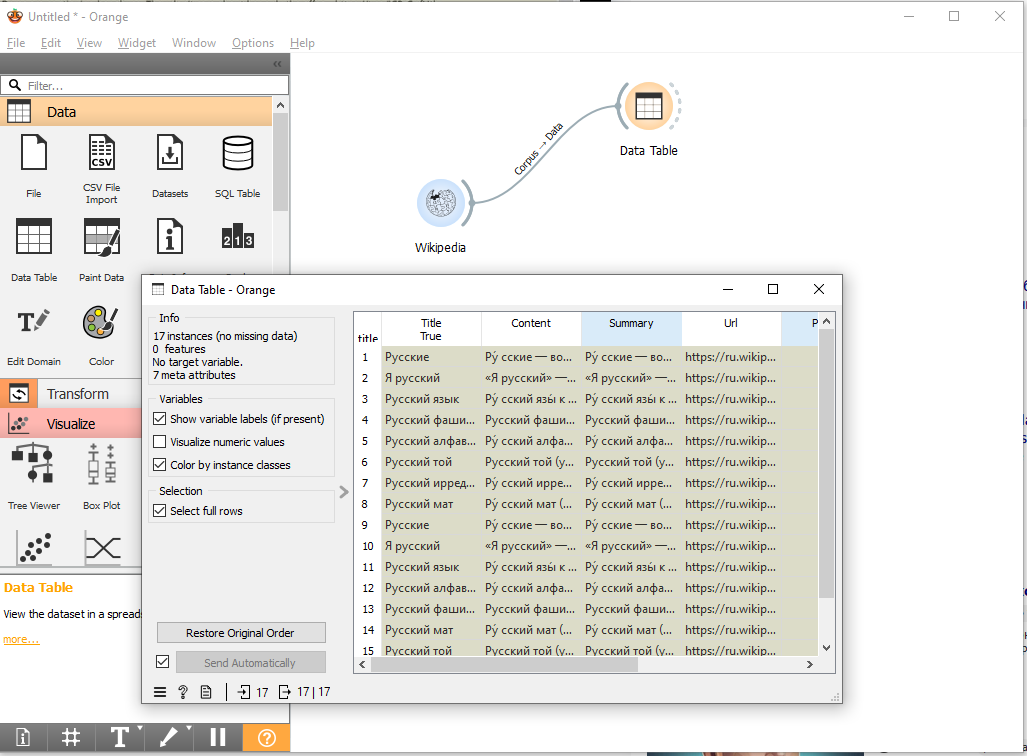
## Amazon Comprehend

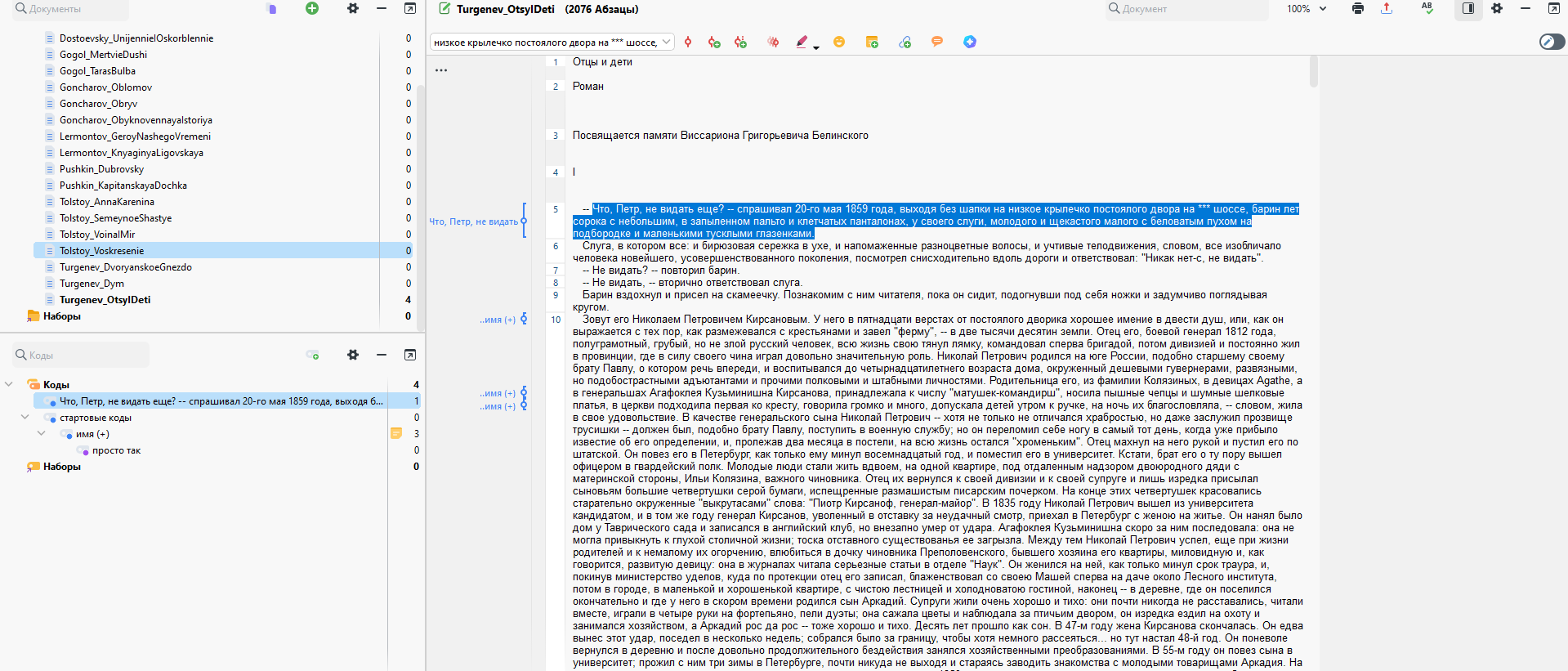
## Apache OpenNLP

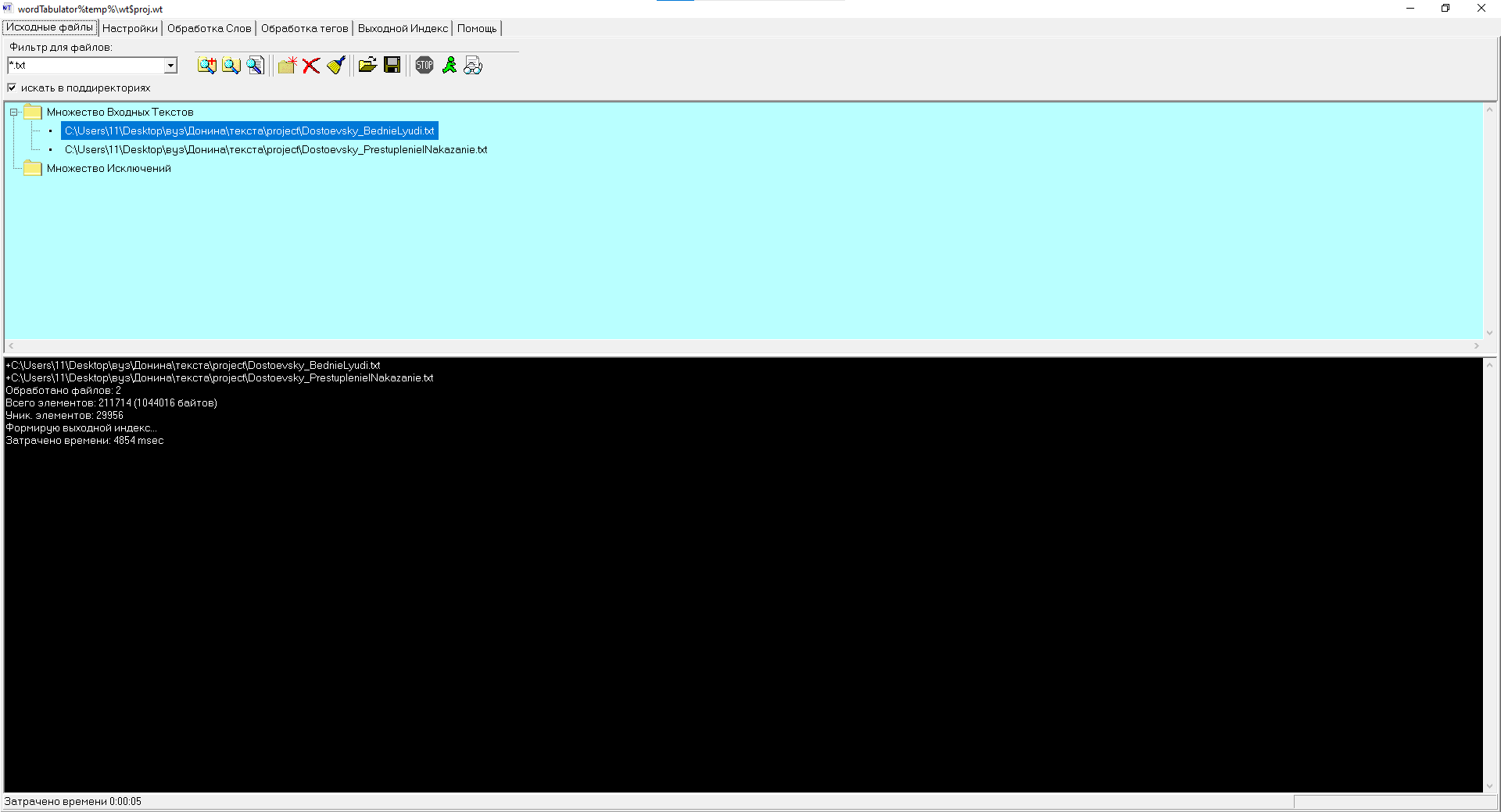
## Aylien

1. Natasha

| Ресурс | Orange | MAXQDA | Visual text | WordTabulator |
| --- | --- | --- | --- | --- |
| удобство интерфейса | + | + | - | + |
| богатый функционал | + | + | - | - |
| Узкая специализация | - | - | + | + |
| можно ли создать таблицу частот? | + | + | + | + |
| анализ частоты слов | + | + | + | + |
| Поддержка скриптов | + | + | + | + |

****





**файл “Анализаторы текстов”**

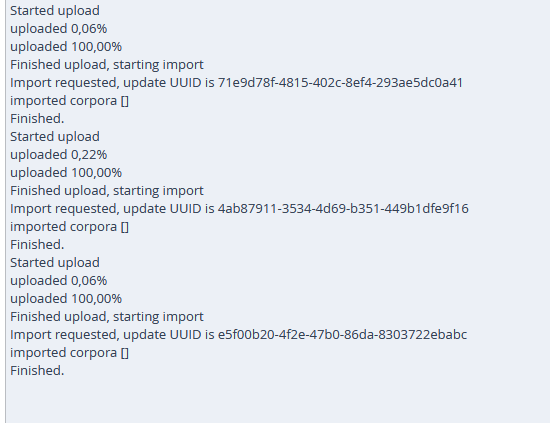
<http://dirtdirectory.org/categories/text-mining> - не работает((

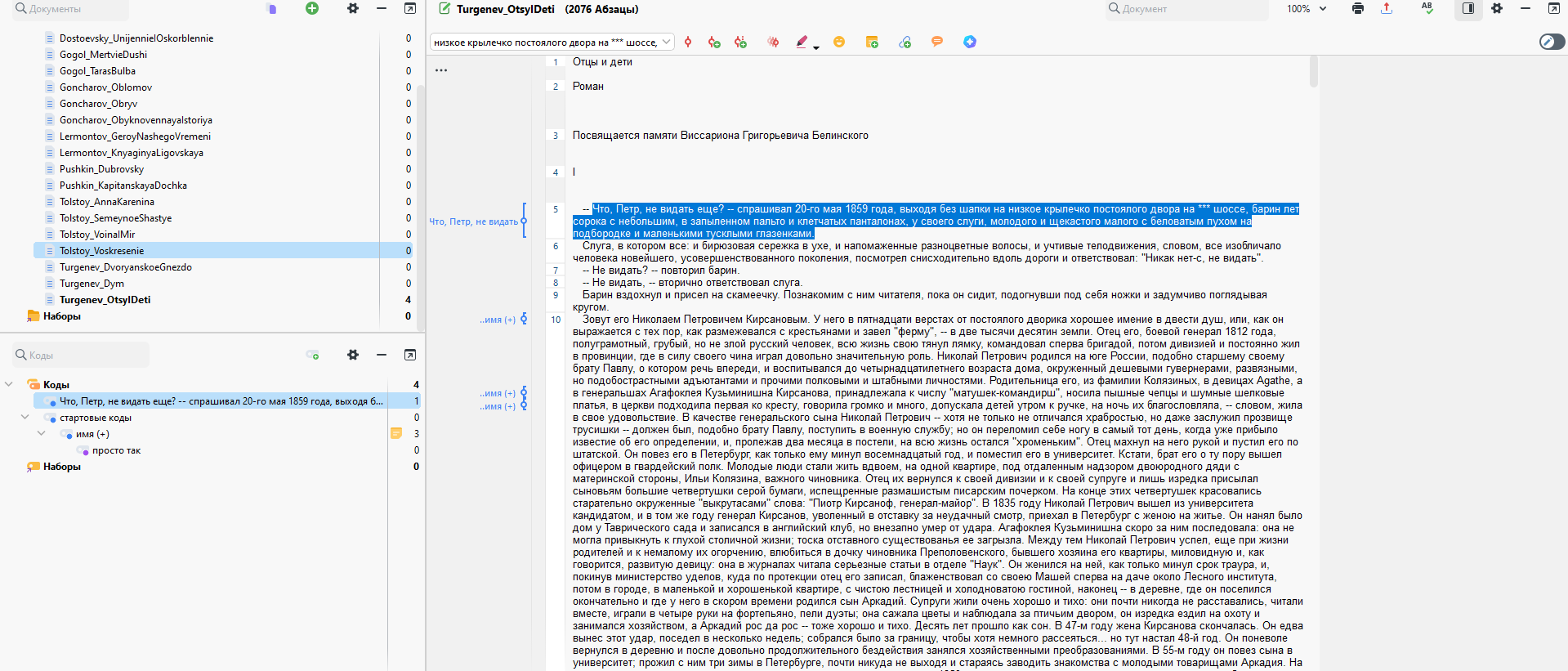
<https://www.textanalyzer.ru/> - ну как работать с букмекерами я понял

аналогов именно онлайн не нашёл

ANNIS - позволяет производить поиск по текстам, статистический анализ, использовать регулярные выражения, поддерживает разные языки, работает с разными форматами текстов, визуализировать свои результаты

Я пытался его запустить, но он каждый раз не загружал корпус, снизу доказательства



MAXQDA - Создание и применение кодов и категорий для систематизации данных, автоматическое кодирование с помощью машинного обучения, создание тематических карт, диаграмм, таблиц и графиков экспорт результатов анализа в различные форматы, генерация отчетов и презентаций для представления данных и многое другое

DiscoverText - Загрузка и импорт текстовых документов из различных источников, Предварительная обработка текста, включая удаление стоп-слов и знаков пунктуации, Токенизация, лемматизация и стемминг для нормализации текста .Частотный анализ для определения наиболее часто встречающихся слов и фраз

Я несколько раз регистрировался на пробный доступ, но мне ни разу не прислали(

Visual Text - с ним другая проблема, об этом ниже. Фунции:

\*Загрузка данных: Импортируйте текстовые документы из различных источников, таких как файлы TXT, CSV и XML.

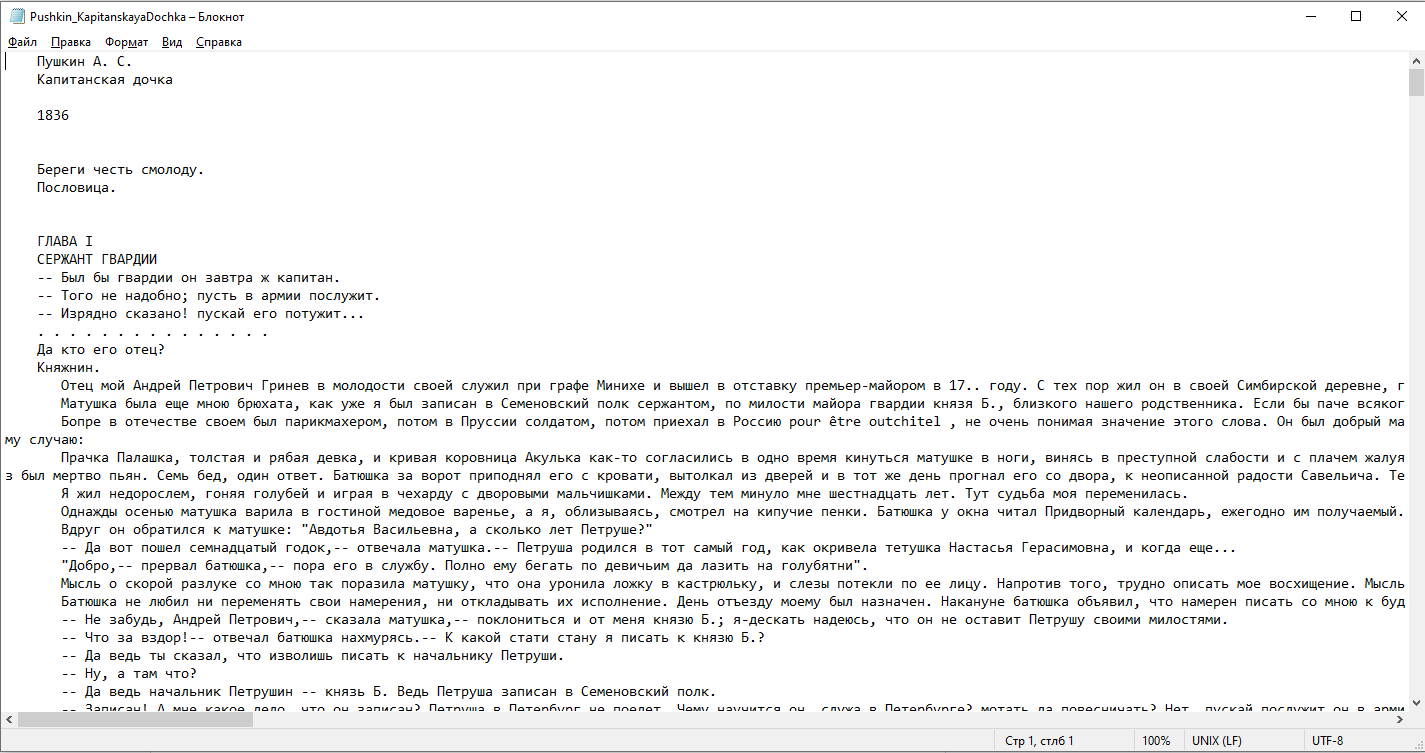
\* Предварительная обработка текста: Очищайте данные, удаляя стоп-слова, знаки пунктуации и другие нежелательные элементы.

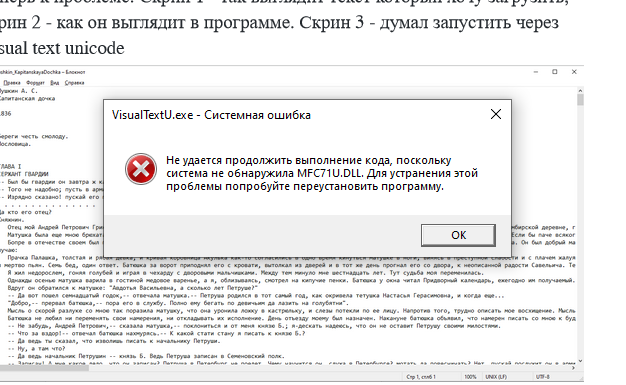
\* Токенизация: Разбивайте текст на токены (отдельные слова или фразы) для последующего анализа.

\* Лемматизация и стемминг: Преобразуйте слова в их основные формы для улучшения точности анализа.

\* Частотный анализ: Определяйте частоту появления слов или фраз в тексте.

\* Построение облаков слов: Визуализируйте частоту появления слов в виде облака слов.

Теперь к проблеме: Скрин 1 - так выглядит текст который хочу загрузить, скрин 2 - как он выглядит в программе. Скрин 3 - думал запустить через Visual text unicode, а он всё время с такой ошибкой



**Уровень 2**

**файл “Разметка и TEI “**

Будет лежать рядом

## **Автоматический анализ текста. Автоматический синтез текста (уровни 2, 3)**

1.

| Слово | Пр. рус. | Пр. англ | комментарий |
| --- | --- | --- | --- |
| сущ. | слон | elephant | на русском предлагает переводы на английский и падежи, на английском - значения |
| гл. | спорить | argue | на русском предлагает переводы на английский и морфологические особенности на английском - значения |
| прилаг. | красивый | beautiful | на русском предлагает переводы на английский и морфологические особенности на английском - значения |
| нар. | тяжко | sore | на русском предлагает переводы на английский и морфологические особенности на английском - значения |

2.

| предложение на русском | линг инф в синт. анализе |
| --- | --- |
| 1 солнце светит |  |
| 2 солнце светит, когда наступает утро |  |
| 3 солнце светит и поют птицы |  |
| Общий комментарий: данные системы анализа можно применять для распознавания связи между словами в предложениях. | |

3. Не работает

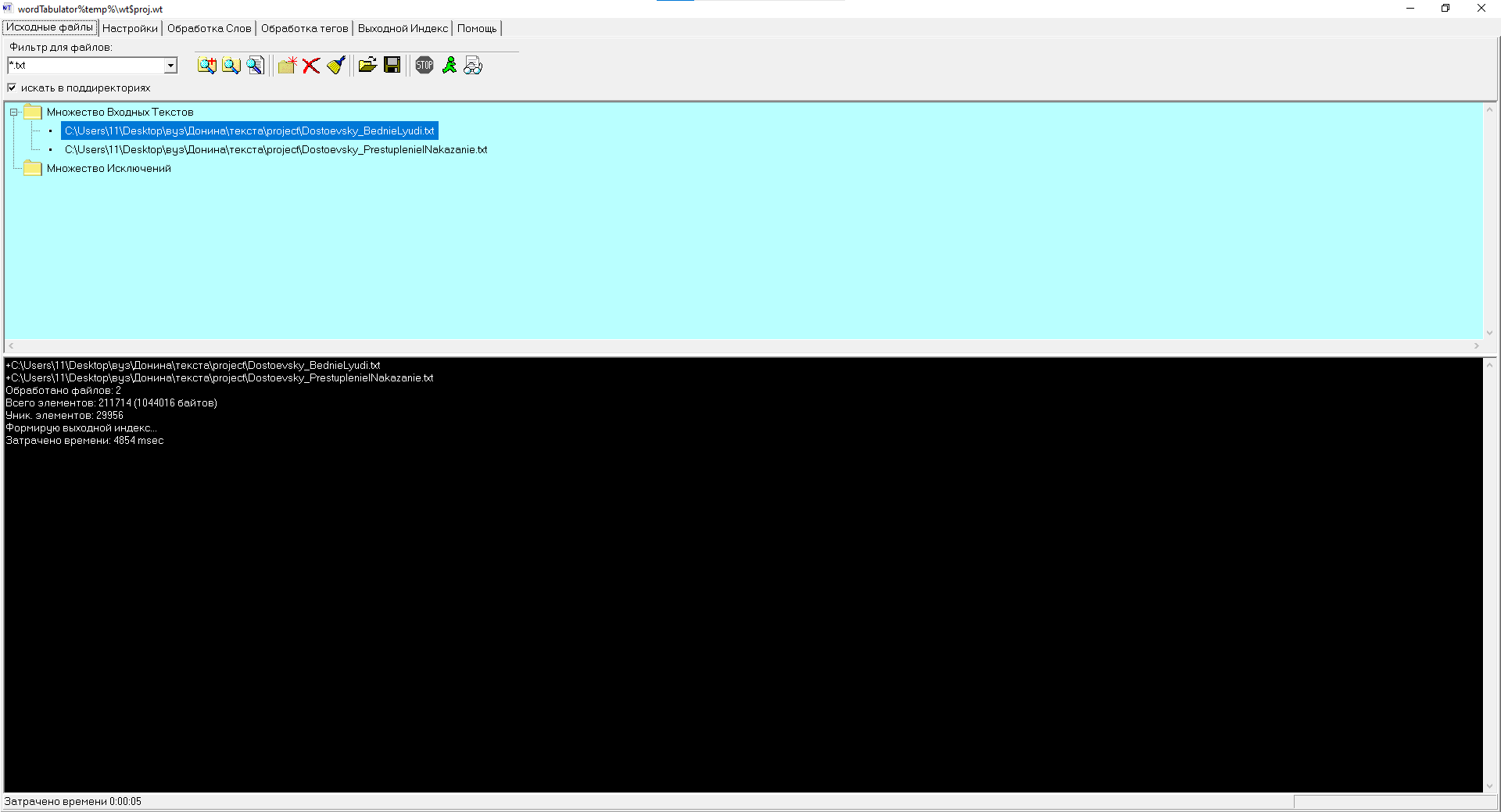
4. Не работает

5.

| Результат: всё отлично |
| --- |
| Комментарий: полезно при написании курсовых и дипломных работы, т.к. позволяет узнать процент |

**файл"Проектное задание 8 Прикладные программные средства"**

В качестве исходных текстов я взял два произведения Достоевского: “Бедные люди” и “Преступление и наказание”



Анализ частотности в двух соседних html файлах.

В текстах Достоевского явно прослеживается, что для него в первую очередь важны ментальная сфера и эмоции. Такие темы не могли не отразиться на тексте: в работах прослеживаются особые лексические конструкции относящиеся, например к “людям из подполья”, также им присуще семантика неопределённости. В своих работах Достоевский поднимает множество самых разных тем, исследует социальные и политические проблемы своего времени, включая критику радикальных идеологий и размышления о роли религии в обществе.

**Уровень 3**

**файл “регулярные выражения”**

исходная:

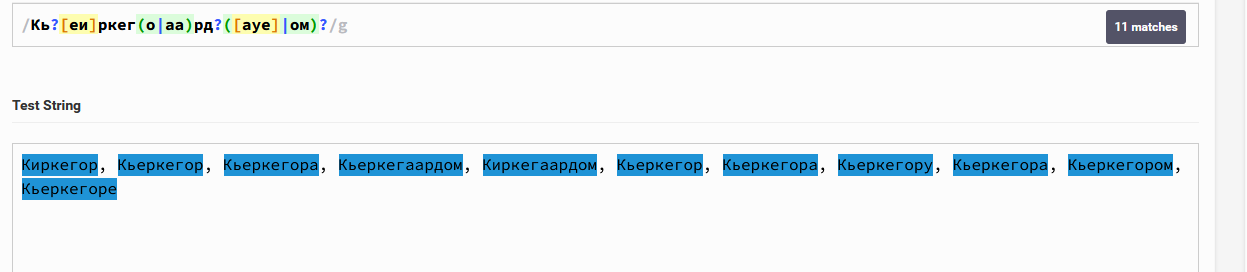
Кь?(е|и)ркег(о|аа)рд?[^\s\.,\?!:;-]{0,2}

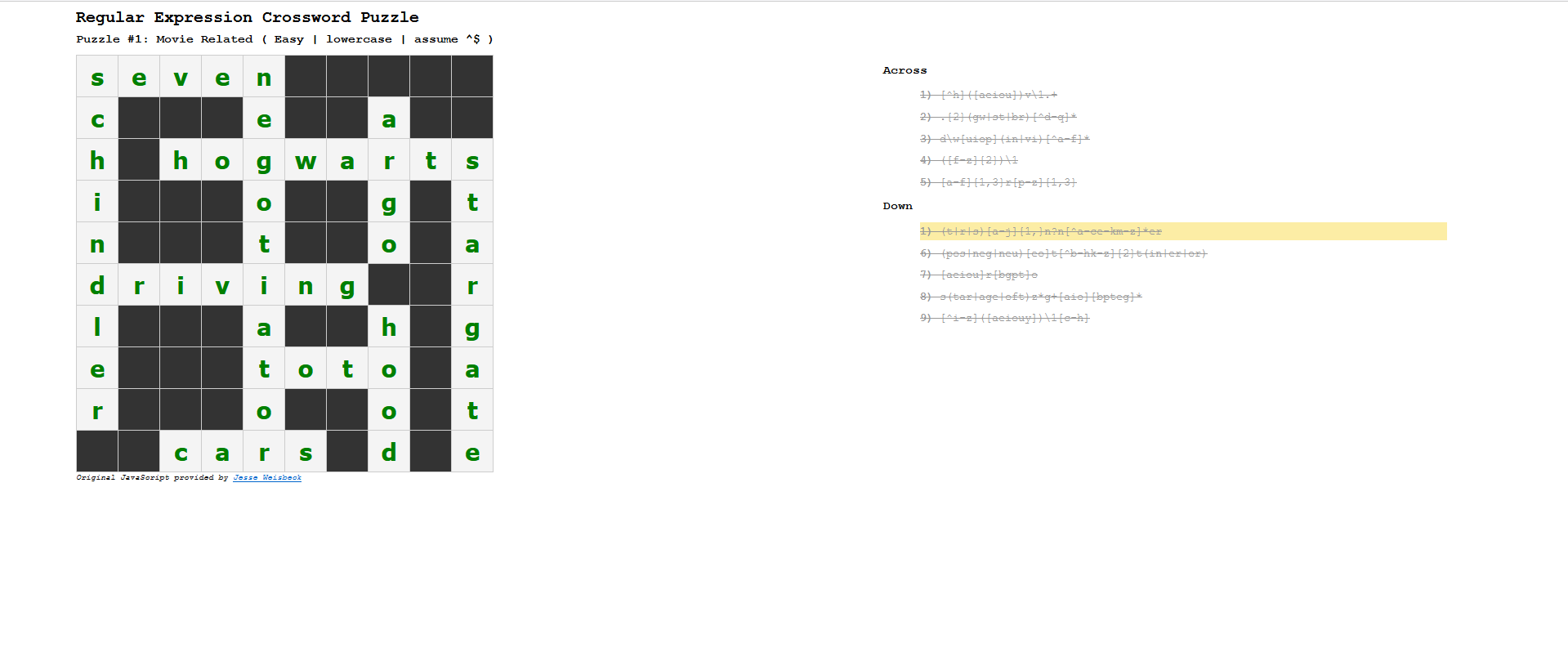
упрощенная версия:

Кь?[еи]ркег(о|аа)рд?(а|е|у|ом)? или Кь?[еи]ркег(о|аа)рд?([ауе]|ом)?

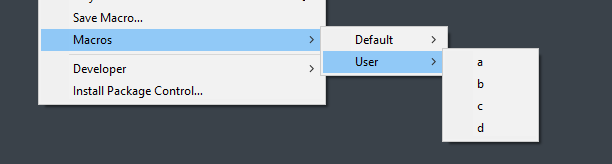
Это работает, но мне кажется это не совсем корректный метод:

.ь?[еи]....(о|аа).д?([ауе]|ом)?

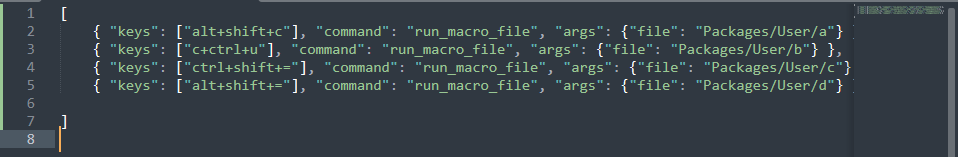


Кроссворд:

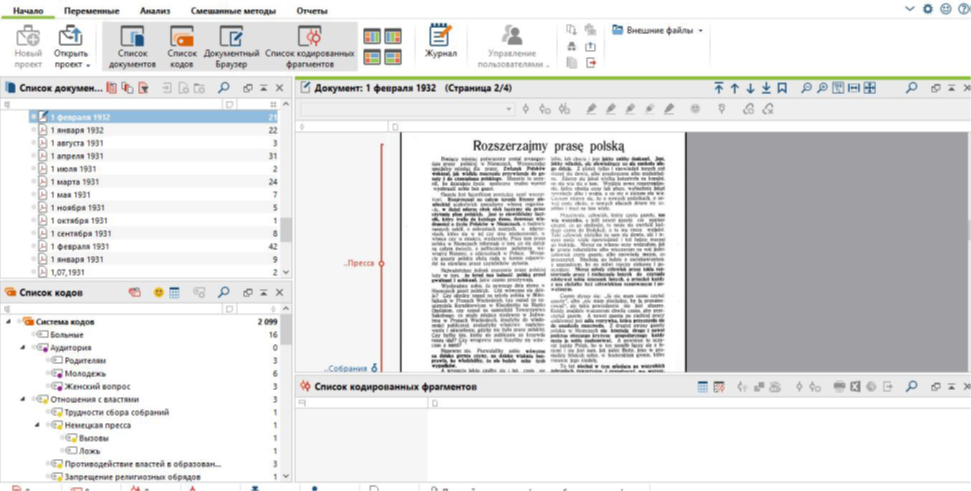
**файл “Форматы данных и кодировки. Макросы в Sublime**

****

****

****

**файл “Анализаторы текстов”**

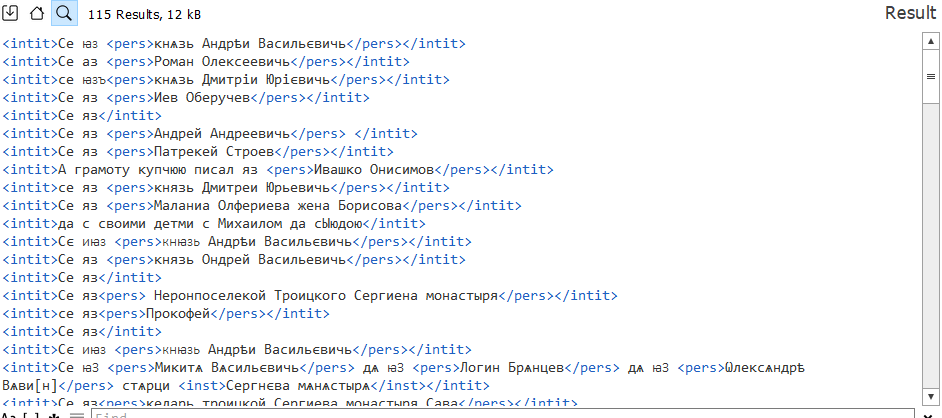
На скриншотах представлена большая коллекция документов 1930-1932 гг в программе MAXQDA. Также кодами размечены документы, среди них можно выделить: Большие, Аудитория(Родители, Молодёжь, Женский вопрос), Отношения с властями(Трудности сбора собраний, Немецкая пресса(Вызовы, Ложь) Противодействие властей в образовании чего-то, Запрещение религиозных обрядов, Борьба), Формы Национальной жизни(Хроника польской жизни, Религия(Праздники, Обряды))

**файл"Лингвистический анализ"**

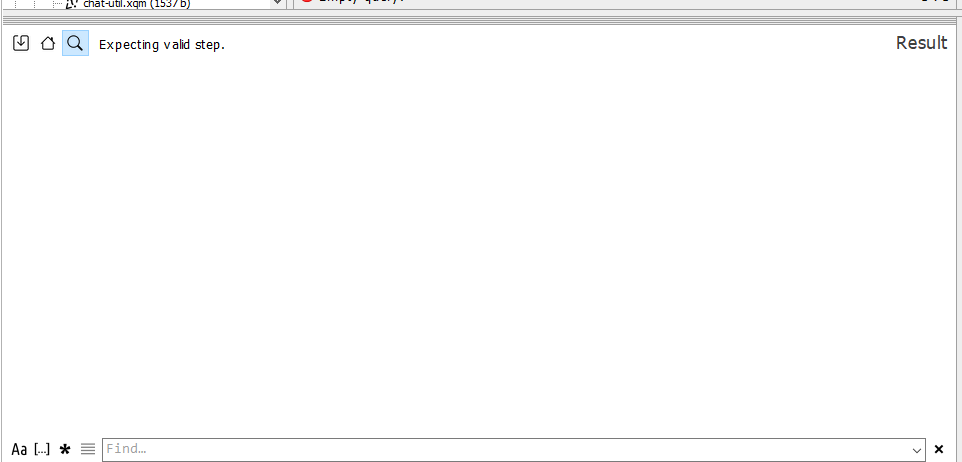
Честно тестировал, хотел ещё питонячьей библиотекой воспользоваться, но не пока не дошли руки.

**файл “BaseX и язык запросов XPath”**

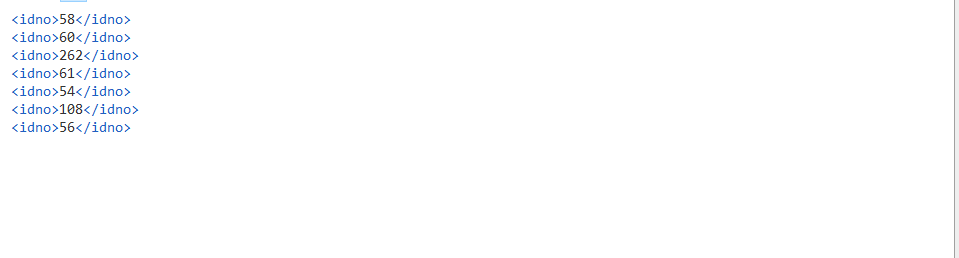
Вот запрос /TEI//intit

****

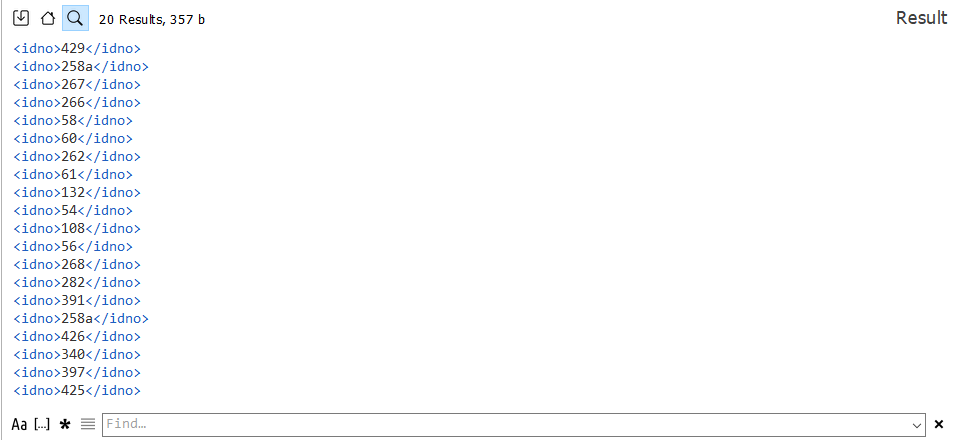
по запросу /TEI//

****

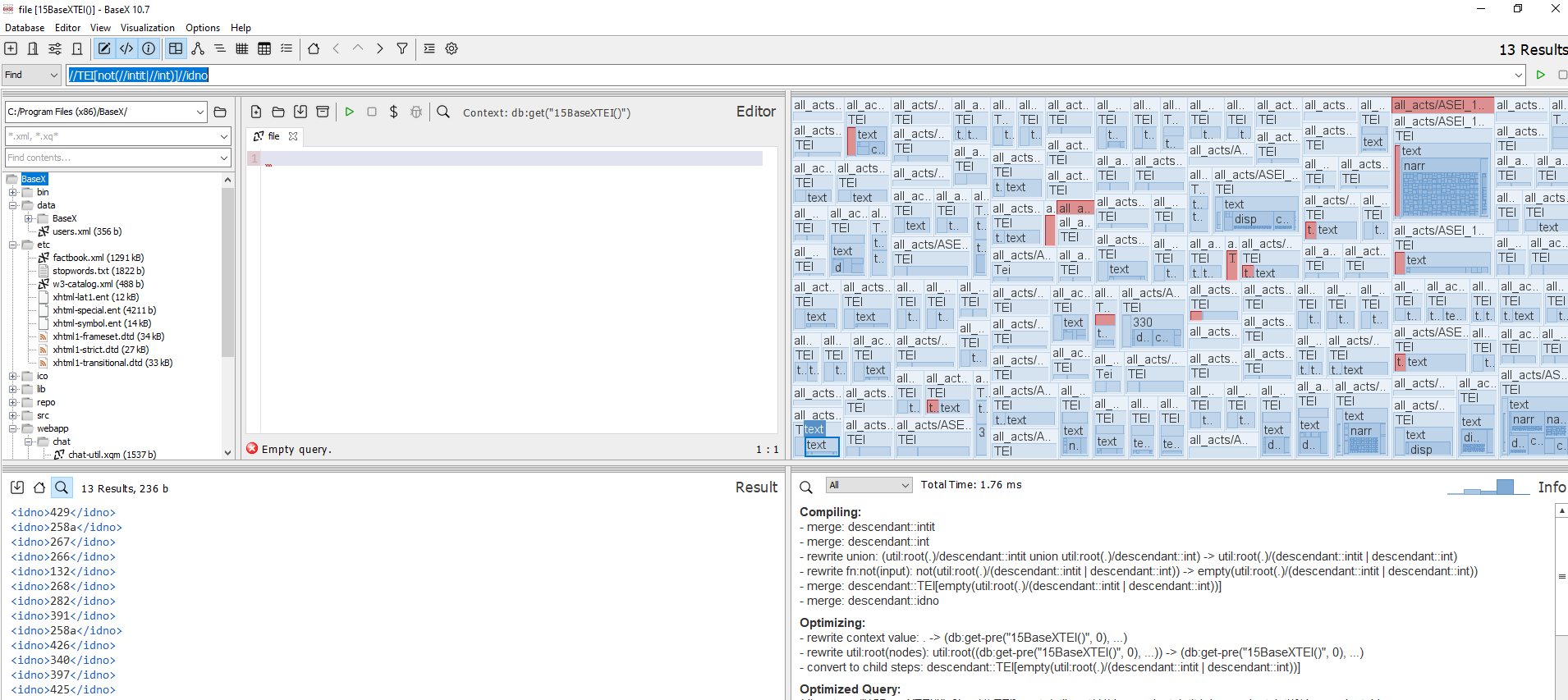
файлы, где есть int но нет intit

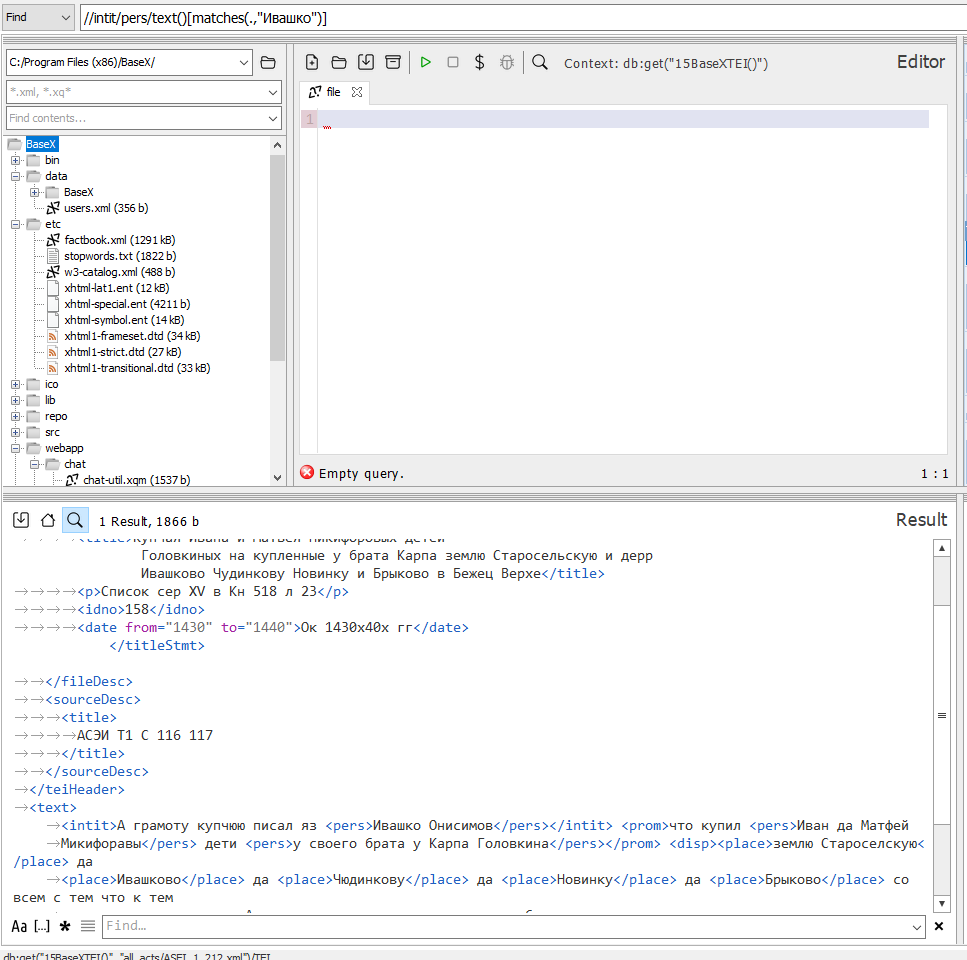
****

файлы, которые не размечаны



где нет не int не intit



грамота с Ивашко - 158

Разнообразие всех видов грамот можно получить с помощью //title/text()[matches(.,"ая")]

## **Автоматический анализ и синтез текста**

| слово | машинная основа | слово | машинная основа |
| --- | --- | --- | --- |
| семья | семь | окружать | окружа |
| весна | весн | сидеть | сиде |
| дело | дел | друг | друг |
| пианино | пианино | ты | т |
| стремительный | стремительн | жёлтый | желт |
| развивать | развива | петь | пе |

1. Кон - конь и коньки

Ноч - ночь и ночник

1. Не понятно какие выделены,если речь про “косой”: там где-то коса, косой, косить
2. рядом лежит
3. приобретать - начинать иметь что-то, утрачивать - переставать иметь что-то, сохранять - не переставать иметь что-то, лишать -кто-то заставляет перестать иметь что-то
4. Мне кажется, что фразы Б сформулированы машиной, т.к. первая сказанная фраза выглядит оторванной от контекста,