Казанский федеральный университет
Институт вычислительной математики и информационных технологий
Кафедра информационных систем

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

# Поисковый веб-сервис для образовательного учебного курса

Обучающийся 4 курса, группы 09-962
Морозов С.А.

Руководитель: канд. техн. наук, доцент
Невзорова О.А.

Казань – 2023

# Актуальность

- Развитие компьютерных технологий предоставляет большие возможности для модернизации процесса обучения, в том числе для разработки  обучающих систем.
- В составе обучающих систем важно иметь поисковый модуль, который выполняет роль справочной системы, позволяющей быстро найти нужный термин и его определение.
- При построении поискового модуля необходимо решить задачу поиска и классификации определений терминов. Данная задача относится к классу задач NER (named entities recognition) (распознавание именованных сущностей).

# Цель работы

- Разработка веб-сервиса для извлечения и поиска определений математических терминов

# Задачи

1. Создать коллекцию математических учебных текстов на основе учебника по геометрии для средней школы и сайта для онлайн обучения yaklass.ru (различные форматы документов: html, txt, xml). Преобразовать к удобному для анализа формату.
2. Разработать синтаксические шаблоны для выделения из предложений математических терминов.
3. Реализовать автоматическое и полуавтоматическое выделение терминов с помощью специального модуля с графическим интерфейсом.
4. Разработать БД для хранения выделенных объектов.
5. Разработать поисковый индекс и интерфейс поисковой системы по математическим терминам.

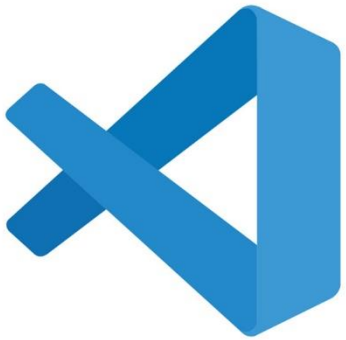
# Используемые ресурсы и технологии

.[RegEx]*

PYMORPHY 3

TKinter


NLTK
Natural Language Toolkit




Beautifulsoup

# Образовательный ресурс YaKlass (https://www.yaklass.ru)

**Ломаной** называется фигура, которая состоит из точек и соединяющих их отрезков.

Точки называются **вершинами** ломаной, а отрезки — **звеньями** ломаной.

Ломаная называется **замкнутой**, если у неё концы совпадают.

Если концы ломаной не совпадают, то она называется **незамкнутой**.

**Многоугольник** — это простая замкнутая ломаная линия и конечная часть плоскости, которую она ограничивает.

Вершины ломаной линии называются **вершинами** многоугольника, а её звенья — **сторонами** многоугольника.

Отрезок, соединяющий две вершины, не лежащие на одной стороне, называется **диагональю** многоугольника.

# Алгоритм парсинга yaklass.ru

1. Парсинг ссылок на уроки с теорией по геометрии (получаем json-файл с ссылками на уроки и названиями на уроки)
2. Парсинг содержимого страницы с теорией, без блоков для навигации по сайту (получаем html-страницы)
3. Парсинг текста из html-страниц (получаем текст без тэгов внутри)

# Фрагмент xml учебника “Геометрия” для 7-9 классов Атанасяна

```
<paragraph>
  <p_id>2</p_id>
  <p_name>Площади параллелограмма, треугольника и трапеции</p_name>
  <section>
    <s_id>51</s_id>
    <s_name>Площадь параллелограмма</s_name>
    <section_text>
      <text>Условимся одну из сторон параллелограмма называть основанием, а перпендикуляр, проведенный из любой точки противоположной стороны к прямой, содержащей ос
      <theorem name="Площадь параллелограмма">
        <t_text>Площадь параллелограмма равна произведению его основания на высоту.</t_text>
        <proof>Рассмотрим параллелограмм ABCD с площадью S. Примем сторону AD за основание и проведем высоты BH и CK (рис.182). Докажем, что S = AD*BH. Докажем снача
        параллелограмма ABCD и треугольника DCK. С другой стороны, она составлена из прямоугольника HBCK и треугольника ABH. Но прямоугольные треугольники DCK и ABH
        стороны параллелограмма, а углы 1 и 2 равны как соответственные углы при пересечении параллельных прямых AB и CD секущей AD), поэтому их площади равны. Следо
        прямоугольника HBCK равна S. По теореме о площади прямоугольника S = BC * BH, а так как BC=AD, то S=AD*BH. Теорема доказана.</proof>
      </theorem>
      <pictures>
        <picture>
          <img src="182.png"/>
          <text>Рис.182</text>
        </picture>
      </pictures>
    </section_text>
  </section>
  <section>
    <s_id>52</s_id>
    <s_name>Площадь треугольника</s_name>
    <section_text>
      <definition name="Основание треугольника">Одну из сторон треугольника часто называют его основанием.</definition>
      <definition name="Высота треугольника">Если основание выбрано, то под словом «высота» подразумевают высоту треугольника, проведенную к основанию.</definition>
```

# Фрагмент DTD файла

!DOCTYPE book [

<!ELEMENT book (book_name,(chapter+))>

<!ELEMENT book_name (#PCDATA)>

<!ELEMENT chapter (chapter_id|chapter_name|paragraph)*>

<!ELEMENT chapter_id (#PCDATA)>

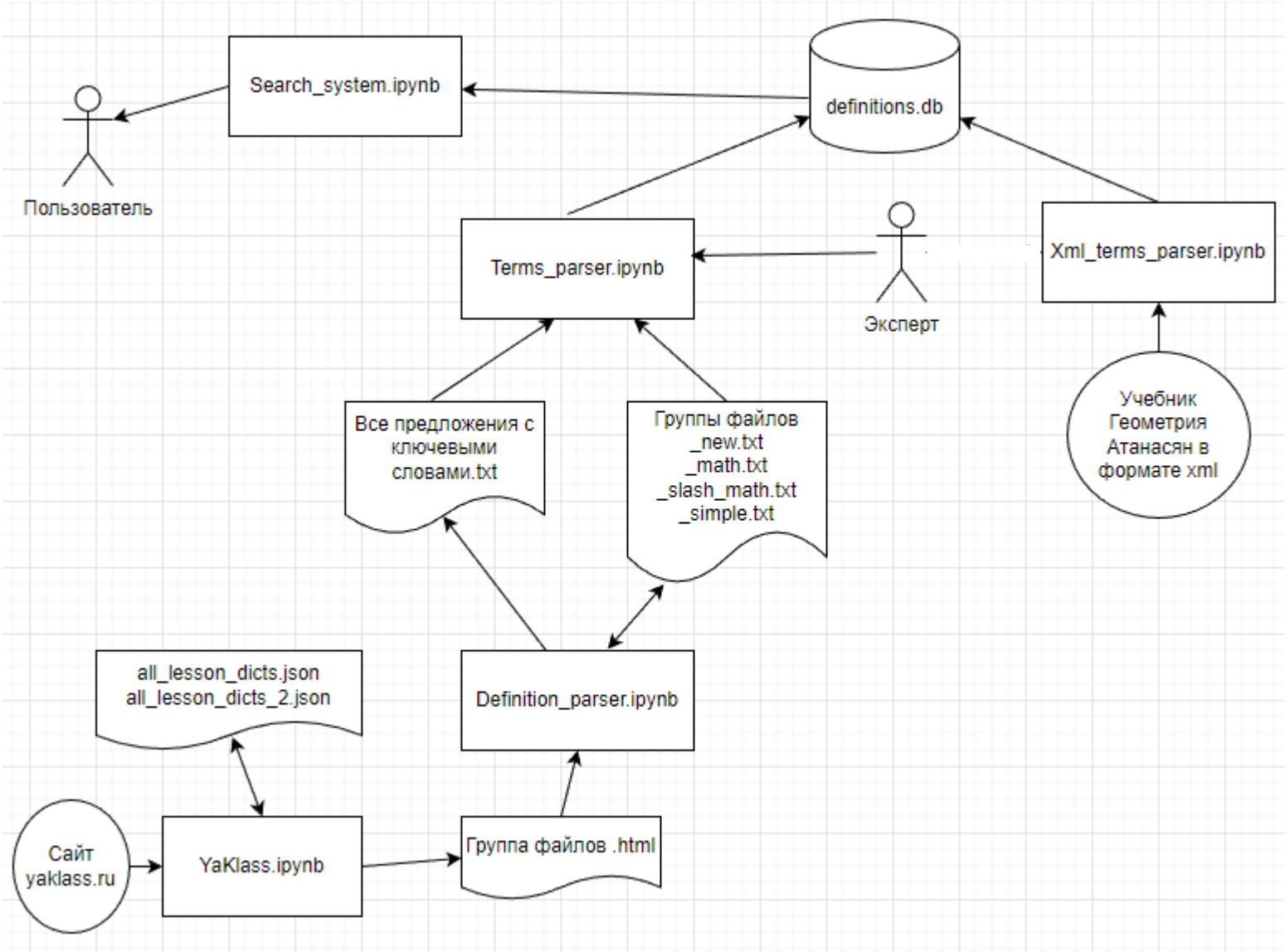
<!ELEMENT chapter_name (#PCDATA)>

<!ELEMENT paragraph (p_id|p_name|section|exercise|exercises_block|questions_block|pictures|exercises|questions|additional_exercises|additional_questions|additional_exercise|s_name|building_tasks|question|hard_exercise|name|picture)*>

<!ELEMENT definition (#PCDATA)>

<!ATTLIST definition name CDATA #REQUIRED>

# АРХИТЕКТУРА программного модуля



# ИЗВЛЕЧЕНИЕ определений и определяемых понятий: подходы

- Лексико-синтаксические шаблоны правил
- Модели машинного обучения (SVM(метод опорных векторов), CRF(метод условных случайных полей))
- Нейронные сети

# Типы определений: примеры

- Определение с текстом

*Угол — геометрическая фигура, которая состоит из точки и двух лучей, исходящих из этой точки.*

- Определение с рисунком

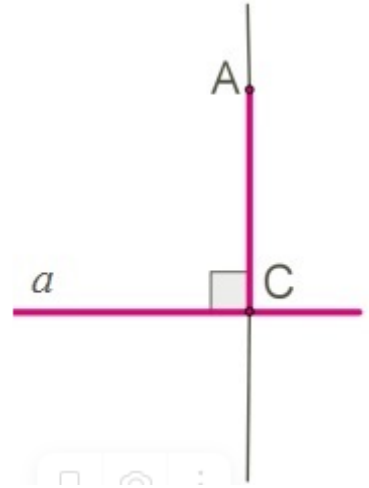
*Отрезок AC называется перпендикуляром, проведенным из точки A прямой a, если прямые AC и a перпендикулярны. Точка C называется основанием перпендикуляра.*

- Определение с математической формулой

**На плоскости две прямые $a$ и $b$, которые не пересекаются, называются параллельными и обозначаются $a \parallel b$.**

Произведением вектора $\vec{a}$ на число $k$ ($k \neq 0$) называется вектор $\vec{b}$, модуль которого равен $\left|\vec{b}\right| = |k| \cdot \left|\vec{a}\right|$, при этом:

- векторы $\vec{a}$ и $\vec{b}$ сонаправлены, если $k > 0$;
- векторы $\vec{a}$ и $\vec{b}$ противоположно направлены, если $k < 0$.



# ИЗВЛЕЧЕНИЕ определений и определяемых понятий: трудности

- *Гомотетия — это преобразование подобия. Это преобразование, в котором получаются подобные фигуры (фигуры, у которых соответствующие углы равны и стороны пропорциональны).*
- *Четырёхугольник, все вершины которого лежат на окружности, называется вписанным в эту окружность, а окружность называется описанной около четырёхугольника.*
- *Если все стороны четырёхугольника касаются окружности, то он называется четырёхугольником, описанным около этой окружности, а окружность — вписанной в четырёхугольник.*
- *Две плоскости называются параллельными, если они не имеют общих точек.*
- *Ломаной называется фигура, которая состоит из точек и соединяющих их отрезков. Точки называются вершинами ломаной, а отрезки — звеньями ломаной.*

# Описание метода извлечения определений

- Метод синтаксических шаблонов (для поиска определений-предложений)
- Метод выделения определяемого термина с помощью шаблона (внутри найденного предложения)
- Примеры шаблонов

| | |
|---|---|
| • NP 1 – <это> NP 1 | Многоугольник — это простая замкнутая ломаная линия и конечная часть плоскости, которую она ограничивает. |
| • NP 1 <называется> NP 5 | Если прямая имеет две общие точки с окружностью, то она называется секущей. |
| • <аналогично определяется> NP 1<br>• <назовем> | Два отрезка называются параллельными, если они лежат на параллельных прямых.<br>Аналогично определяется параллельность отрезка и прямой, отрезка и луча, двух лучей, луча и прямой.<br>… Назовем его единичным полукругом. |
| • NP 1 <будем называть> NP 5 . | Синусом острого угла прямоугольного треугольника будем называть отношение противолежащего катета к гипотенузе. |
| • NP1 <называемый> NP5 | Многогранник, называемый призмой, можно построить следующим образом. |
| • NP5 <называется> NP1 | Трапецией называется четырехугольник … |
| • NP 1 <является> ADJ | Если …, то треугольник является прямоугольным. |

# Анализ синтаксических шаблонов

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| В | этом | разложении | коэффициенты | координатных | векторов | называют | координатами | вектора | m | . | | | | |
| None | loct | loct | nomn | gent | gent | None | ablt | gent | None | None | | | | |
| PREP | NPRO | NOUN | NOUN | ADJF | NOUN | VERB | NOUN | NOUN | None | None | | | | |
| | | | | | | | | | | | | | | |
| Координатами | вектора | являются | координаты | конечной | точки | этого | вектора | , | если | вектор | располож | так | , | чт |
| ablt | gent | None | nomn | gent | gent | gent | gent | None | None | nomn | None | None | None | N |
| NOUN | NOUN | VERB | NOUN | ADJF | NOUN | NPRO | NOUN | None | CONJ | NOUN | PRTS | CONJ | None | C |
| | | | | | | | | | | | | | | |
| Четырёхугольн | , | все | вершины | которого | лежат | на | окружности | , | называет | вписанны | в | эту | окружнос | , |
| nomn | None | None | gent | gent | None | None | gent | None | None | ablt | None | accs | nomn | N |
| NOUN | None | PRCL | NOUN | ADJF | VERB | PREP | NOUN | None | VERB | PRTF | PREP | ADJF | NOUN | N |
| | | | | | | | | | | | | | | |
| Если | все | стороны | четырёхугольник | касаются | окружнос | , | то | он | называет | четырёху | , | описанным | около | эт |
| None | None | gent | gent | None | gent | None | None | nomn | None | ablt | None | ablt | None | ge |
| CONJ | PRCL | NOUN | NOUN | VERB | NOUN | None | CONJ | NPRO | VERB | NOUN | None | PRTF | PREP | AD |
| | | | | | | | | | | | | | | |
| Окружность | называют | описанной | около | треугольника | , | если | все | вершины | треугольн | располож | на | окружности | . | |
| nomn | None | gent | None | gent | None | None | None | gent | gent | None | None | gent | None | |
| NOUN | VERB | PRTF | PREP | NOUN | None | CONJ | PRCL | NOUN | NOUN | PRTS | PREP | NOUN | None | |
| | | | | | | | | | | | | | | |
| Окружность | называют | вписанной | в | треугольник | , | если | все | стороны | треугольн | касаются | окружнос | . | | |
| nomn | None | gent | None | nomn | None | None | None | gent | gent | None | gent | None | | |
| NOUN | VERB | PRTF | PREP | NOUN | None | CONJ | PRCL | NOUN | NOUN | VERB | NOUN | None | | |

# Шаблоны

| Часть | плоскости | , | ограниченная | окружностью | , | называется | кругом | . |
|---|---|---|---|---|---|---|---|---|
| nomn | loct | None | nomn | ablt | None | None | None | None |
| NOUN | NOUN | None | ADJF | NOUN | None | VERB | ADVB | None |

28 определение

| Если | две | геометрические | фигуры | удаётся | совместить | наложением | , | они | — | равные | . |
|---|---|---|---|---|---|---|---|---|---|---|---|
| None | accs | nomn | nomn | None | None | ablt | None | nomn | None | accs | None |
| CONJ | NUMR | ADJF | NOUN | VERB | INFN | NOUN | None | NPRO | None | ADJF | None |

4 определения

| Угол | — | геометрическая | фигура | , | которая | состоит | из | точки | и | двух | лучей | , | исходящих | из | этой | точки | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| accs | None | nomn | nomn | None | nomn | None | None | gent | None | gent | gent | None | gent | None | gent | gent | None |
| NOUN | None | ADJF | NOUN | None | ADJF | VERB | PREP | NOUN | CONJ | NUMR | NOUN | None | ADJF | PREP | ADJF | NOUN | None |

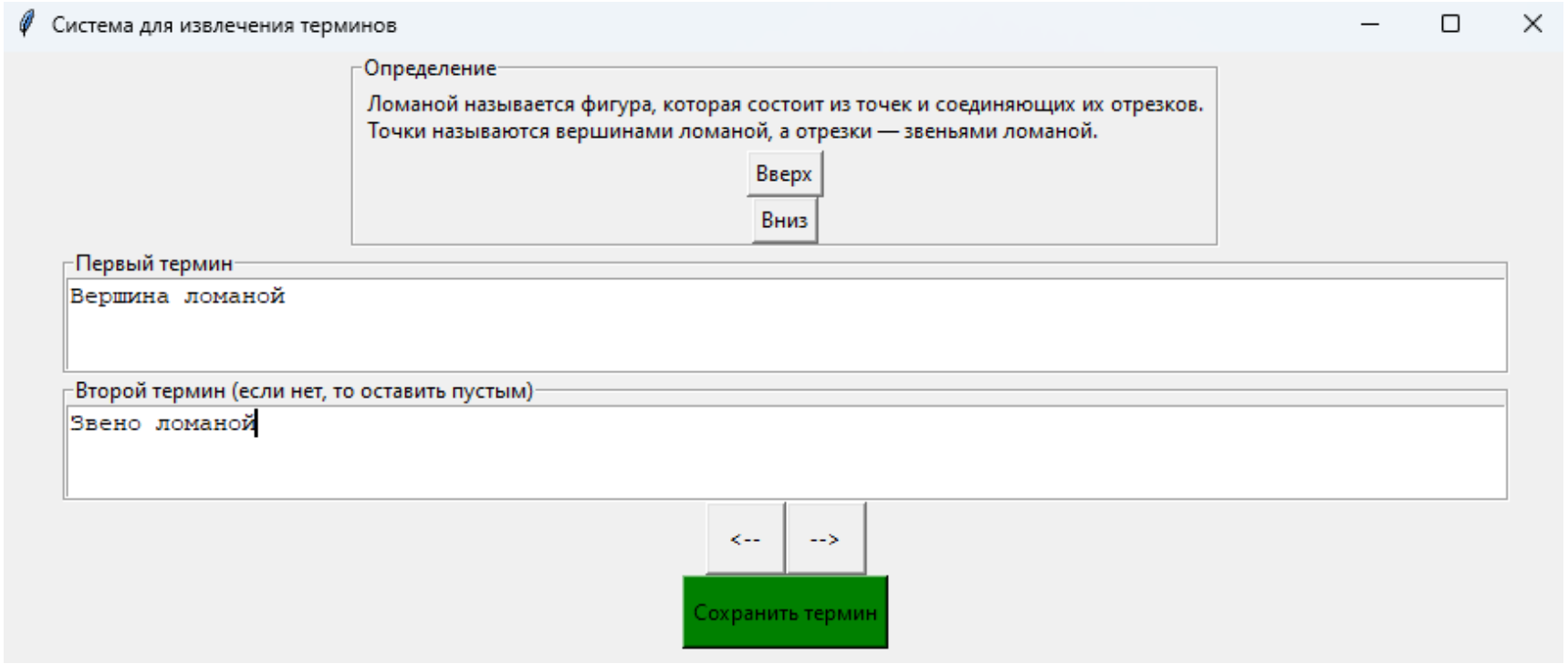
33 определения

# Результаты автоматического извлечения определений

- Сколько определений
- Автоматическое 33 штуки
- Полуавтоматическое

| Определения (всего) | 207 |
|---|---|
| Определения (yaklass.ru) | 146 |
| Опр.  по учебнику геометрии Атанасяна | 61 |
| Опр. $NP^1$ <— это> $NP^1$ | 33 |
| “,” + VERB(н-р, “является”) + $NP^5$ | 28 |
| Опр. $NP^1$ + VERB(н-р, “называют”) + $NP^5$ (и перевернутая формула) | 42 |

# Модуль извлечения определений (полуавтоматический)



# Организация хранения математических определений

| | termid | firstterm | secondterm | definition | document |
|---|---|---|---|---|---|
| 55 | 55 | Параллельный перенос фигуры | {null} | Параллельным переносом фигуры называется перенос всех точек пространства на одно расстояние в одном направлении. | YaKlass.ru |
| 56 | 56 | Сторона, противолежащая углу | Угол, противолежащий стороне | Сторону, которая лежит напротив угла, называют противолежащей углу, и угол называют противолежащим стороне. | YaKlass.ru |
| 57 | 57 | Прилежащие углы | {null} | Углы, которые имеет одну общую сторону, называют прилежащими этой стороне. | YaKlass.ru |
| 58 | 58 | Периметр треугольника | {null} | Сумма сторон треугольника называется периметром . | YaKlass.ru |
| 59 | 59 | Смежные углы | {null} | Два угла, у которых одна сторона общая, а две другие являются продолжением одна другой, называются с межными . | YaKlass.ru |
| 60 | 60 | Вертикальные углы | {null} | Два угла называются вертикальными , если обе стороны одного угла являются продолжениями сторон другого. | YaKlass.ru |
| 61 | 61 | Перпендикулярные прямые | {null} | Если две пересекающиеся прямые образуют четыре прямых угла, они называются перпендикулярными . | YaKlass.ru |
| 62 | 62 | Равновеликие многоугольники | {null} | Если многоугольники имеют равные площади, но они не равные, то их называют равновеликими . | YaKlass.ru |
| 63 | 63 | Подобные треугольники | {null} | Два треугольника называются подобными , если их соответствующие углы равны, а соответствующие стороны пропорциональны. | YaKlass.ru |

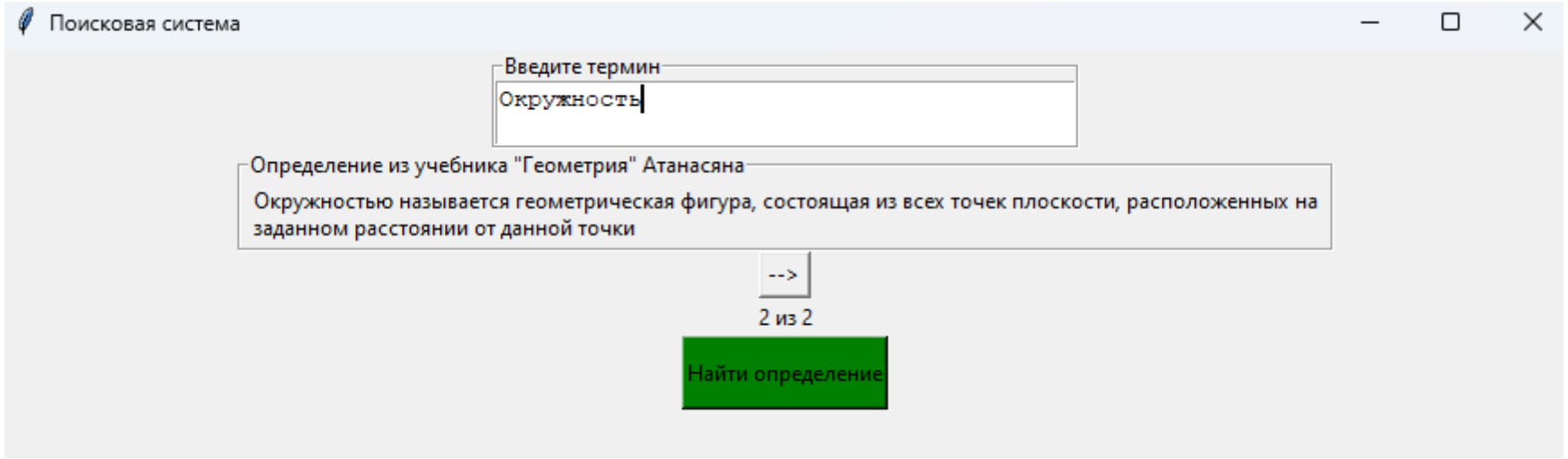
146 дефиниционных предложений извлечено с сайта ЯКласс

# Организация хранения математических определений

| | | | | |
|---|---|---|---|---|
| 86 | Окружность | {null} | Окружностью называется геометрическая фигура, состоящая из всех точек плоскости, расположенных на заданном расстоянии от данной точки | Атанасян... |
| 87 | Условие и заключение теоремы | {null} | Во всякой теореме различают две части: условие и заключение. Условие теоремы - это то, что дано, а заключение - то, что требуется доказать. Рассмотрим, например, теорему, выражающую признак параллельности двух прямых: если при пересечении двух прямых секущей накрест лежащие углы равны, то прямые параллельны. В этой теореме условием является первая часть утверждения: «При пересечении двух прямых секущей накрест лежащие углы равны» (это дано), а заключением - вторая часть: «Прямые параллельны (это требуется доказать). | Атанасян... |
| 88 | Теорема обратная данной | {null} | Теоремой, обратной данной, называется такая теорема, в которой условием является заключение данной теоремы, а заключением - условие данной теоремы. | Атанасян... |
| 89 | Внешний угол | {null} | Внешним углом треугольника называется угол, смежный с каким-нибудь углом этого треугольника. | Атанасян... |

61 дефиниционное предложение извлечено из учебника «Геометрия» Атанасяна

# Построение поисковых  запросов к базе данных



# Заключение

1. Создана коллекцию математических учебных текстов. Преобразована к удобному для анализа формату.
2. Разработаны синтаксические шаблоны для выделения из предложений математических терминов.
3. Реализовано автоматическое и полуавтоматическое выделение терминов с помощью специального модуля с графическим интерфейсом.
4. Разработана БД для хранения выделенных объектов.
5. Разработан поисковый индекс и интерфейс поисковой системы по математическим терминам.

Спасибо за внимание!