Метод обработки естественного языка в условиях слабой структурированности исходного текста

Кобук М.Г.
Атаева О.М.
Тучкова Н.П.
Теймуразов К.Б.

03 декабря 2025

План доклада

- Введение
- Цель и область исследования
- Решение::Транслятор
- Решение::Поиск
- Дальнейшие шаги

Введение

Базы знаний

- Единое представление
- Структурированный формат
- Детерминированный подход
- Автоматизация

Входные данные

- Разное представление
- Разные типы
- Большие объёмы

Обработка

- Онтология
- Обучение моделей
- Заполнение каталогов
- Полнотекстовый поиск

Требования

- Быстро
- Качественно
- Легковесно

Исходные данные

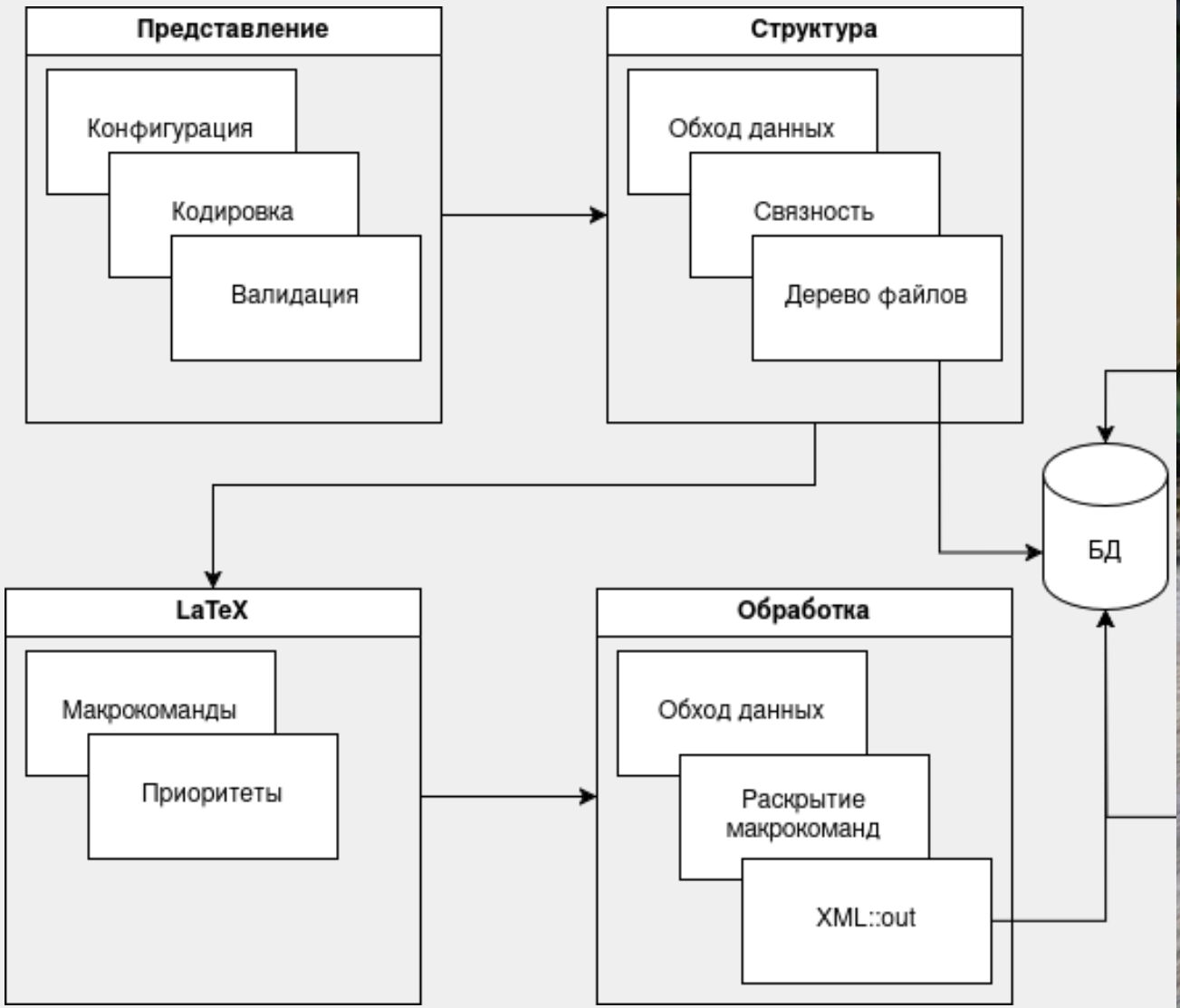
- В рамках проекта SciLibRu
- Выпуски периодических изданий
  - Многофайловый состав
  - Пользовательские макрокоманды
  - Нестабильный формат
  - Опора на редкие/устаревшие библиотеки
  - Нестандартные кодировки
- Отдельно стоящие файлы

! Нужно универсальное, быстро адаптируемое решение

Сравнение функций

	tex2xml	LaTeXML	TeX4ht	t2x3
Сложные файлы	−	+	+	+
Многофайловые док-ты	−	−	−	+
Простая конфигурация	*	−	+/−	+/−
Пользовательские макро	−	+	+/−	+
Стабильность	+	+	−	−
Строгий шаблон	*	+	+/−	−/+
Нестандартные файлы	−	−	−	+

Описание подхода [Трансляция]



Представление и структура

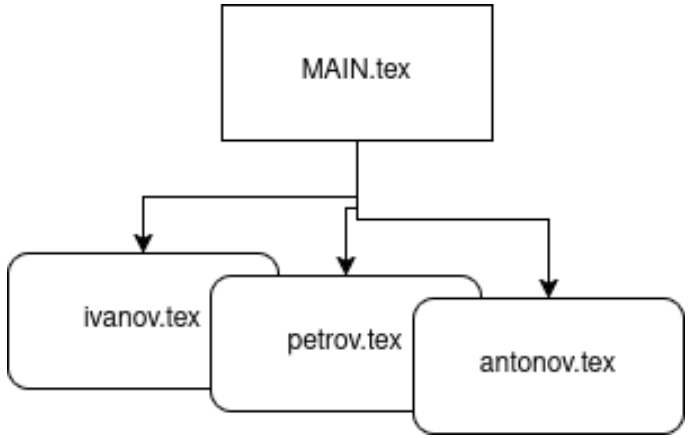
Кодировка

```
□↩秽 □I □≒$\alpha$  \-□-□-□ □˙˜$\gamma$,
 □砩□稽~$X$ 祭. □ □□
□⊑□$a$ ~$b$ ∴, □ $|X|\hm\le a$  $|\gamma|\hm\le b$,
```

Дерево

```
journal/
├── antonov.tex
├── ivanov.tex
├── MAIN.tex
└── petrov.tex
```

⇒



Макрокоманды

- Регистрация макрокоманд в два прохода
- Чёрный список и приоритет макрокоманд
- Итеративное раскрытие макрокоманд

Не всё надо раскрывать

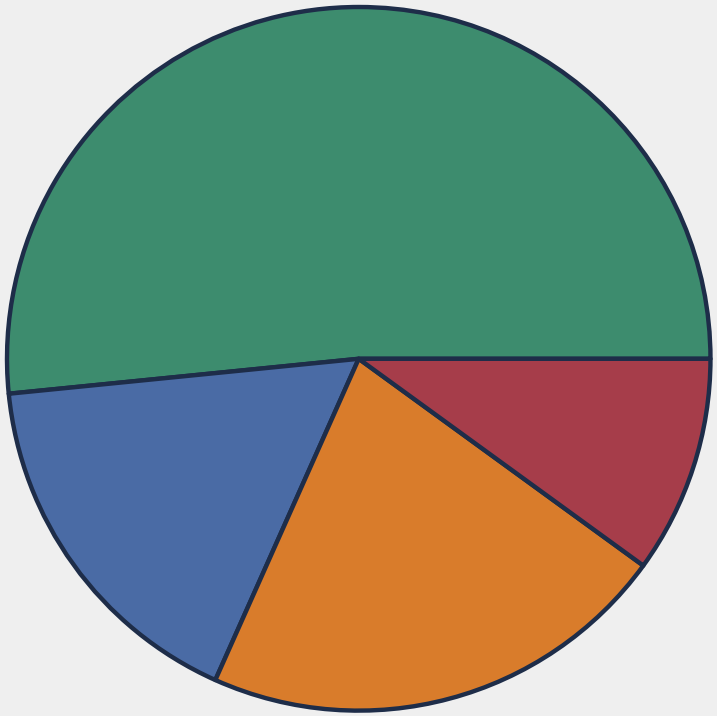
```
\Abst{АННОТАЦИЯ}

\begin{center}\small\nwt
\parbox{150mm}{%\baselineskip=2.5ex
\textbf{Аннотация:}\ \
АННОТАЦИЯ}\end{center}}
```

Макрокоманды

1. Лексер
   - Текст
   - LaTeX-команда
   - Управляющий символ
   - [Формула]
2. Парсер
   - Итеративное раскрытие макрокоманд согласно конфигурации
   - Разные режимы строгости обработки ввода
3. Построение и обход AST
4. Оценка качества (integrity < 0.1; quality > 2.75)

$$integrity = \frac{TeX_{xml}}{TeX_{tex}}; quality = \frac{Heads_{xml}}{Heads_{tex}} + \frac{Tails_{xml}}{Tails_{tex}} + \frac{Formula_{xml}}{Formula_{tex}}$$

Результаты обработки тестовой выборки (1010 файлов)



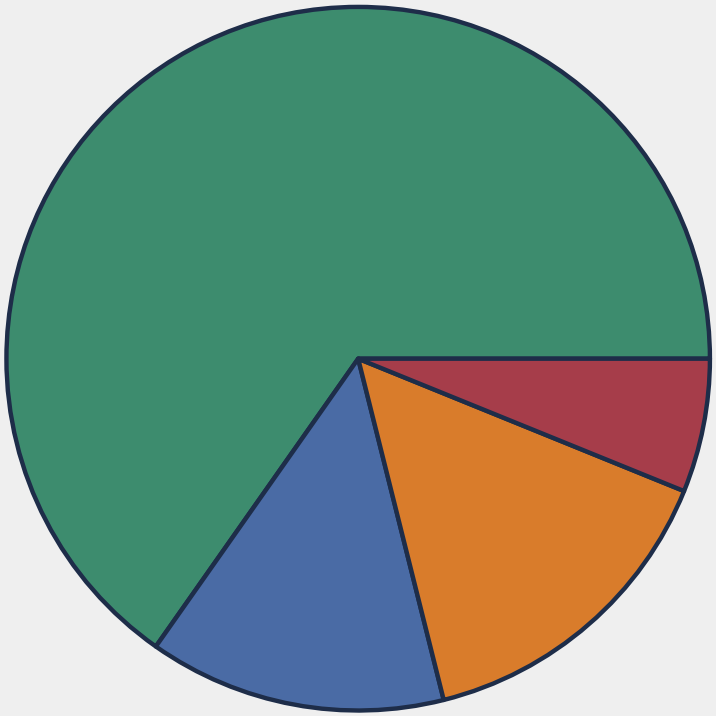
- Без ошибок (51.6%)
- Доработка конфигурации (16.7%)
- Фатальные ошибки (21.7%)
- Расширение покрытия TeX (10%)

- 89 технических/нецелевых файлов исключены из общего числа (1099)
- В ≈12% файлов присутствуют ошибки структуры выходных данных
- В ≈33% файлов наблюдаются остаточные конструкции LaTeX

Улучшения

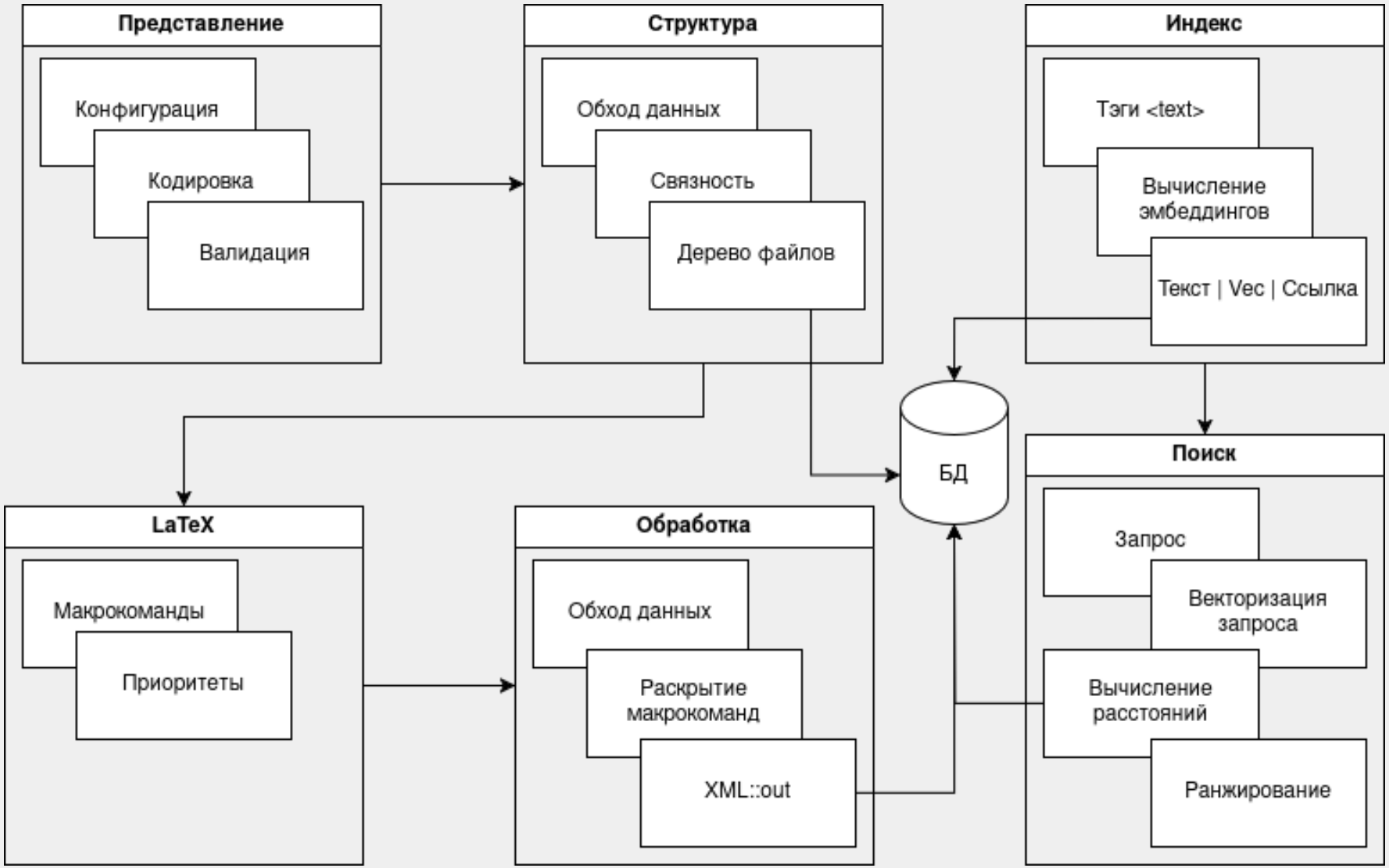
1. Лексер/Парсер
   - Расширение покрытия LaTeX-примитивов и встроенных команд
   - Рекурсивный спуск $\rightarrow$ LALR
   - Автомат состояний
2. Анализ структуры документа
   - Упрощение конфигурации
3. Индикация ошибок
   - Расширение логирования

Результаты доработки



- Без ошибок (65.2%)
- Доработка конфигурации (13.4%)
- Фатальные ошибки (15%)
- Расширение покрытия TeX (6.1%)

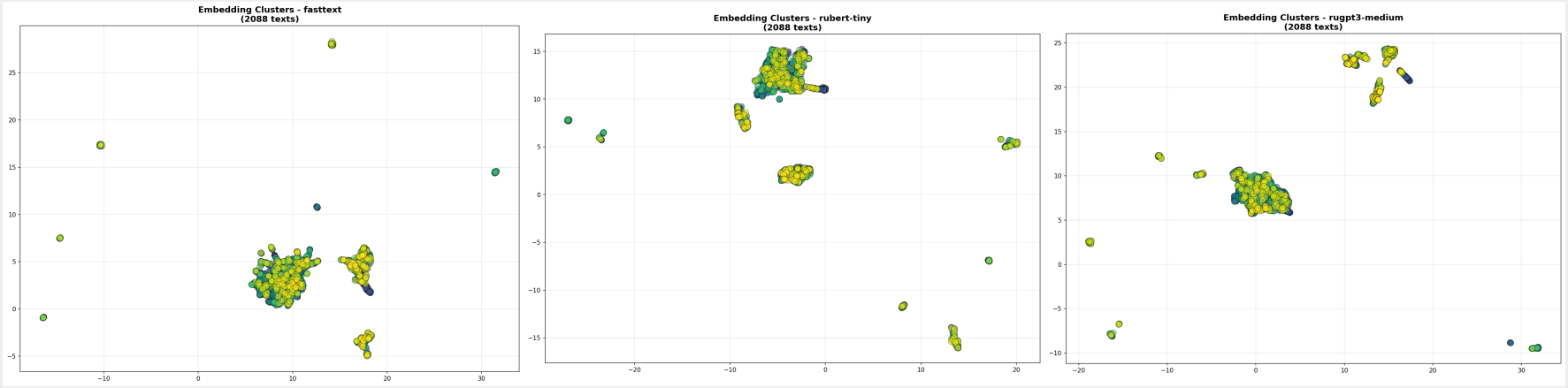
Описание подхода



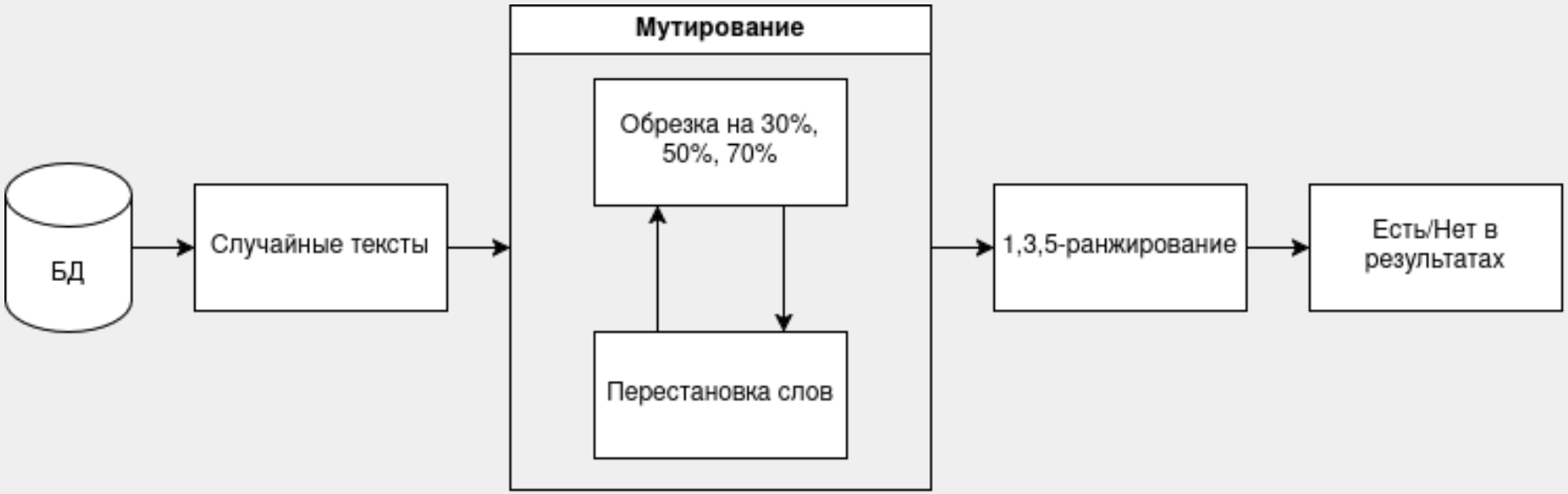
Модели и метрики

- Модели
  - fasttext
  - rubert-tiny
  - rugpt3-medium
  - glove
  - labse
  - ruadapt
  - rubert-base
  - word2vec
  - (tf-idf)
  - (elmo)
- Метрики
  - Косинусное сходство
  - Евклидово расстояние
  - Манхэттенское расстояние
  - Расстояние Пирсона
  - Мера Чебышёва
  - Мера Минковского
  - Мера Махаланобиса

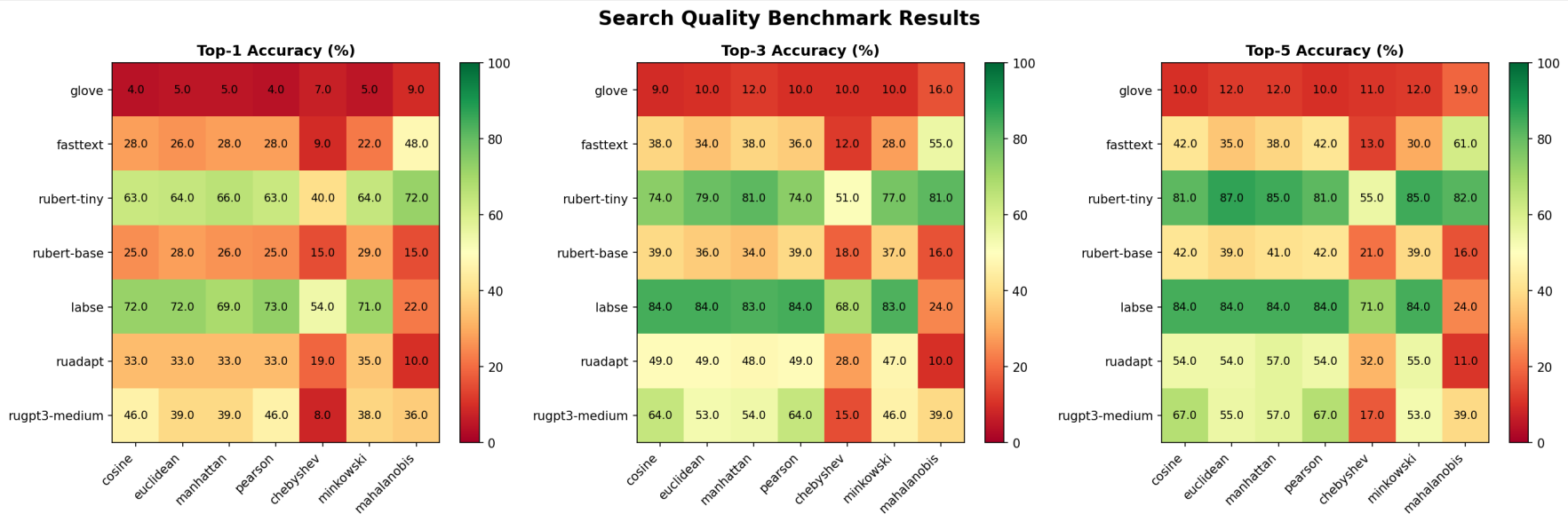
Распределение векторов



Оценка поиска



Оценка поиска



Дальнейшие шаги

1. Расширение покрытия TeX, доработка конфигурации
2. Интеграция с моделями для семантической верификации
3. Дообучение моделей на корпусе для повышения точности поиска
4. Двухуровневая модель поиска
5. Тестирование ELMo

Спасибо за внимание

Лирическое отступление

Лирическое отступление

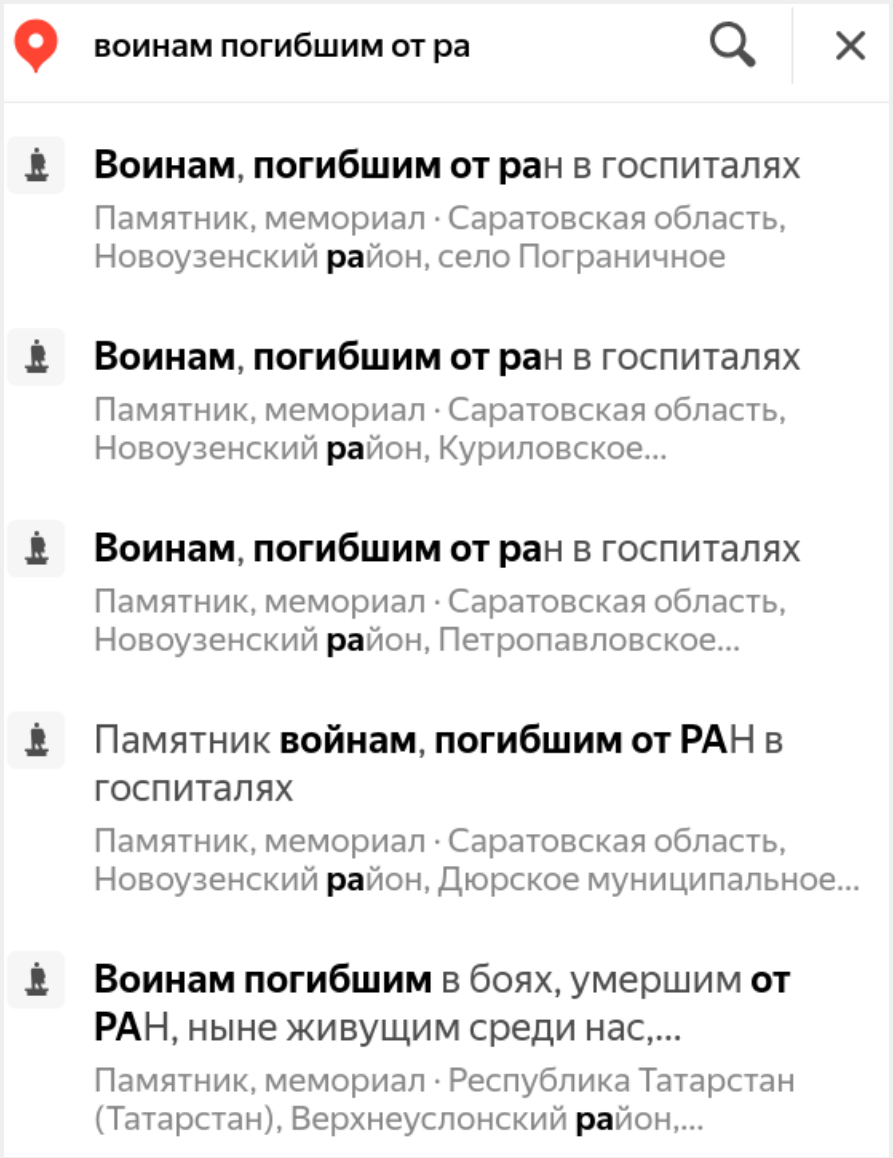
Вот таблица со списком памятников и их наименованиями. Проанализируй таблицу и вставь слово ран/РАН там, где требуется. Будь внимателен к регистру.

Название памятника Местоположение
Кладбище советских воинов, погибших от ? в госпиталях Новосибирска (1941-1945) г. Новосибирск
Мемориал солдатам, погибшим от ? в госпиталях г. Казань

Кладбище советских воинов, погибших от **ран** в госпиталях Новосибирска (1941-1945) г. Новосибирск
Мемориал солдатам, погибшим от **РАН** в госпиталях г. Казань

2 / 2

Лирическое отступление



Объективное отступление

```
Φ(x)=1/√(2π) ∫_{-∞}^{x} e^{-z^2/2} dz, x∈ℝ.
```

```
<formula id="6">\Phi(x)=\displaystyle\frac{\displaystyle 1\mathstrut}
{\displaystyle \sqrt{2\pi}\mathstrut}\int\limits_{-\infty}^{x}e^{-z^2/2}\,dz\,,
\enskip x\in\mathbb{R}\,.</formula>
```

Объективное отступление

```
      <section id="1" name="Introduction. The Weibull distribution">
        <text>
          <value>In probability theory and mathematical statistics, the Weibul
ribution concentrated on the nonnegative half-axis, whose tail decreases in an
          <value>It is named after the Swedish scientist Waloddi Weibull (1887
n the analysis of the strength of materials and described and studied it in de
his distribution in describing many statistical patterns.</value>
          <value>Let γ>0. The Weibull distribution with shape parameter γ is t
          <formula id="1">{\sf P}\left(W_{\gamma}&lt;x\right)= \left[1-e^{-x^{
R}\,.</formula>
```

Объективное отступление

```
Here and далее,
```

Поиск [word2vec+cosine]

- поверхность шар
  1. Таким образом, в однодипольной модели обратная задача решается следующим способом. Нужно найти на поверхности сферы две точки с максимальным значением радиальной компоненты магнитного поля...
  2. т. е. функция ширины экрана непрерывна и в окрестности точки...
  3. меньшим значениям модуля градиента на рис. 5 соответствуют более светлые серые пикселы. Так, при наименьшем значении модуля градиента (...
  4. во всей плоскости...
  5. Расчет количества рабочих мест, которые могут быть размещены в области комфортного наблюдения, должен опираться на способ их размещения. Например, для построения ситуационного зала, где в центре будет находиться овальный стол, расчет будет заключаться в определении максимального размера такого стола...

Поиск [word2vec+mahalanobis]

- Распределение в пределе
  1. используем предел...
  2. распределение Вейбулла называется распределением Рэлея
  3. На основе полученных результатов была написана программа, позволяющая вычислять распределения времен выхода из множества состояний, совместное стационарное распределение числа заявок в системе и состояния системы и связанные с ним характеристики, а также исследовать поведение рассматриваемой СМО в з...
  4. (т. е. «изначально» предельным распределением минимальной порядковой статистики должно быть экспоненциальное), но объем выборки имеет вид...

Поиск [word2vec+minkowski]

- Распределение в пределе
  1. распределение Вейбулла называется распределением Рэлея...
  2. (т. е. «изначально» предельным распределением минимальной порядковой статистики должно быть экспоненциальное), но объем выборки имеет вид...
  3. При этом аналогичная нижняя доверительная граница, вычисляемая указанным выше упрощенным методом (Ллойда и Липова), основанным на непосредственном использовании частных доверительных границ для параметров отдельных элементов, равна