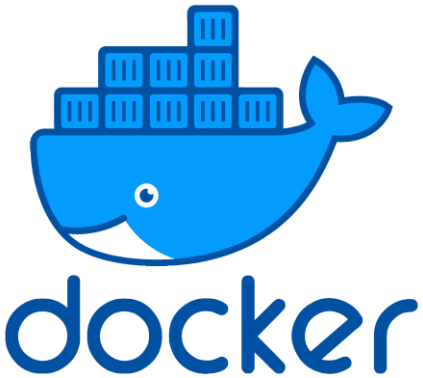
Поисковая система для хранения и суммаризации документов

Функционал

- Индексация новостных статей (Lenta.ru) в OpenSearch
- Поиск документов по текстовому и векторному соответствию
- Векторизация текста с sentence-transformers/all-MiniLM-L6-v2
- Генерация кратких суммаризаций с легковесной LLM (Mistral, gemma и т.д.)
- Веб-интерфейс **OpenSearch Dashboards** для отладки



Сравнение векторного поиска и комбинированного с полнотекстовым

- Комбинированный поиск в целом демонстрирует более качественное нахождение документов с точным совпадением текста и более эффективную обработку заголовков
- В свою очередь чисто векторный поиск не уступает комбинированному с точки зрения нахождения похожих по смыслу документов
- Недостаток реализованного векторного поиска – выдача отдельных нерелевантных документов среди первых, - в дальнейшем может быть решен путем использования более мощной модели векторизации

Характерный пример

Запрос

```
query {
  search(searchString: "Криштиану Роналду перешел из Реала в Ювентус") {
    listDocument {
      id
      title
      text
    }
    total
  }
}
```

Результат векторного поиска


```
"data": {
  "search": {
    "listDocument": [
      {
        "id": "j2WS-pQBP7a3Qw6d9YDL",
        "title": "Семья Роналду нашла лучшего футболиста мира",
        "text": "Катя Авейру, сестра нападающего «Ювентуса» и сборной
         "id": "02WS-pQBP7a3Qw6dLG9S",
         "title": "Сыгравший украинца белорус объяснил ошибку Первого канала",
         "text": "Житель Минска Виталий Юрченко объяснил, почему ему пришлось
          "id": "4WWQ-pQBP7a3Qw6dU0tq",
          "title": "Выплативший 13 миллионов штрафа Роналду снова оказался в суде",
          "text": "Нападающий туринского «Ювентуса» Криштиану Роналду признал себя
          "id": "rWWR-pQBP7a3Qw6dmWOS",
          "title": "Роналду нашел для Месси новый вызов",
          "text": "Нападающий «Ювентуса» Криштиану Роналду заявил,
```


Как видим, второй документ не является релевантным запросу, но в целом точность приемлема

Результат комбинированного поиска


```
"data": {
  "search": {
    "listDocument": [
      {
        "id": "BGWR-pQBP7a3Qw6dSF5x",
        "title": "Тренер «Реала» отреагировал на выпады Роналду",
        "text": "Главный тренер мадридского «Реала» Сантьяго Солари
        "id": "rWWR-pQBP7a3Qw6dmWOS",
        "title": "Роналду нашел для Месси новый вызов",
        "text": "Нападающий «Ювентуса» Криштиану Роналду заявил,
        "id": "B2WS-pQBP7a3Qw6d1X7j",
        "title": "Роналду отказался возвращаться в Мадрид",
        "text": "Нападающий туринского «Ювентуса» Криштиану
        "id": "j2WS-pQBP7a3Qw6d9YDL",
        "title": "Семья Роналду нашла лучшего футболиста мира",
        "text": "Катя Авейру, сестра нападающего «Ювентуса» и с
```


Комбинированный поиск демонстрирует более высокую точность, первый нерелевантный документ представлен лишь на 9 позиции

Сравнение результатов суммаризации с дроблением текста на чанки и без

- Дробление большого текста на чанки помогает произвести более точную суммаризацию, не теряя при этом смысловой составляющей каждого сегмента текста
- Суммаризация реализована «цепочным методом» так, что на каждом чанке обобщение текста производится с учетом предыдущего контекста
- Итоговый метод позволяет получить **краткую** сводку по большому числу документов

Характерный пример

Запрос

```
query {
 summarize(ids: ["32WQ-pQBP7a3Qw6dtlO1",
                 "02WR-pQBP7a3Qw6dc2Av",
                   "M2WQ-pQBP7a3Qw6dY01b",
                   "kGWQ-pQBP7a3Qw6dEkaA",
                   "F2WR-pQBP7a3Qw6dDVrl"])
}
```

Результат суммаризации без дробления на чанки

```
message=Message(role='assistant', content='16 команд уже определены
для плей-офф Лиги Чемпионов сезона 2018/2019. Среди них - Боруссия
(Дортмунд), Атлетико, Барселона, Тоттенхэм, ПСЖ и Ливерпуль (группа
C), Порту и Шальке (группа D). В группе E вышли Бавария и Аякс.
Первые два места в группе F заняли Манчестер Сити и Лион. Мадридский
Реал и Рома прошли из группы G, а туринский Ювентус и Манчестер
Юнайтед - из группы H. В этом сезоне в Лиге Европы представят Россию
два клуба: Зенит и Краснодар. ЦСКА проиграл в финальном матче
группового этапа Лиги Чемпионов мадридскому Реалу со счётом 0:3. Это
позволило испанской команде выйти в плей-офф, а чешской \"Виктории\"
пробиться в Лигу Европы. Жеребьевка 1/16 турнира Лиги Европы
состоится 17 декабря.', images=None, tool_calls=None)"
```

Полученная суммаризация часто бывает несвязной и иногда допускает ошибки (в данном примере неверно приведен результат матча ЦСКА – Реал Мадрид)

Результат суммаризации с дроблением на чанки

```
message=Message(role='assistant', content=\"1. The text discusses the results of the 2018/2019 Champions League and Europa League.\\n   2. In the Champions League, 16 teams have progressed to the knockout stage, including Schalke and Porto from Group D. The final will be held on June 1, 2019, at the Wanda-Metropolitano stadium in Madrid.\\n   3. Notable teams that advanced include Borussia Dortmund, Atletico Madrid, Barcelona, Tottenham Hotspur, Paris Saint-Germain, Liverpool, Porto, Schalke, and Real Madrid.\\n   4. In the Europa League, qualifiers include Rapid Vienna (Austria), Eintracht Frankfurt (Germany), Lazio (Italy), Genk (Belgium), Malmö FF (Sweden), Sevilla (Spain), Krasnodar (Russia), Dynamo Kyiv (Ukraine), Rennes (France), Chelsea (England), and BATE Borisov (Belarus).\\n   5. In Russia, Zénith and Krasnodar will represent in the Europa League for the spring season.\\n   6. The draw for the 1/16 finals of the Europa League will take place on December 17.\\n   7. In a Group G match, Moscow's CSKA defeated Real Madrid at Santiago Bernabeu with a score of 0:3, while Czech team Viktoria Plzen defeated Roma 2:1.\\n   8. As a result, CSKA earned seven points but did not advance to the knockout stage of the Europa League, finishing fourth in its group.\", images=None, tool_calls=None)"
```

Суммаризация с дроблением на чанки демонстрирует более объемный результат и сохраняет ключевые моменты каждого из текстов. Но при этом модель mistral периодически переводит результат на английский язык.