|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ **Информатика и системы управления**

КАФЕДРА **Компьютерные системы и сети (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа, обработки и интерпретации больших данных**

**Отчет**

|  |  |
| --- | --- |
| **по лабораторной работе №** | 5 |

**Название:** Работа с документной NoSQL БД Elasticsearch и реализация процесса data science

**Дисциплина:** Технология параллельных систем баз данных

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ6-12М |  |  | С.В. Астахов | |
|  | (Группа) |  | (Подпись, дата) | | (И.О. Фамилия) |
|  |  |  |  | |  |
| Преподаватель |  |  |  | |  |
|  |  |  | (Подпись, дата) | | (И.О. Фамилия) |

Москва, 2023

**Введение**

**1. Цель работы:** студент должен изучить модель представления данных и способы работы с документной NoSQL базой данных Elasticsearch. Получить навыки инсталляции, индексации и поиска в Elasticsearch, реализовать этапы процесса Data Science, протестировать работу с большим объёмом данных.

**Ход выполнения**

**2. Установка и запуск ES.**

Для упрощения установки воспользуемся докер-контейнером с ES, пробросив порт 9200 в сеть основной машины (рисунок 1).

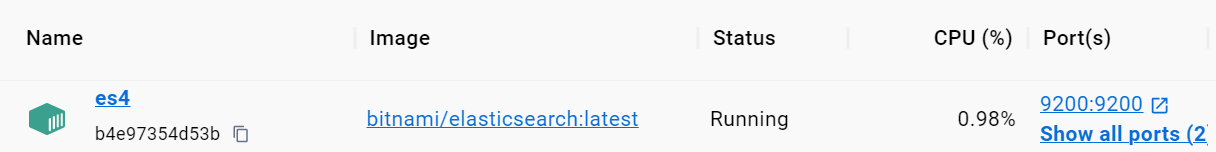


Рисунок 1 — контейнер с ES

Проверим работу ES с помощью GET-запроса к url “http://localhost:9200/?pretty” (рисунок 2).

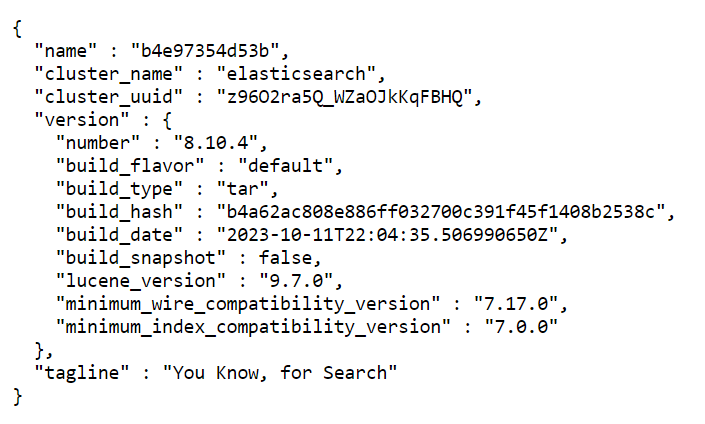


Рисунок 2 — информация о кластере ES

Так же установим пакет для работы с ES в python (рисунок 3).

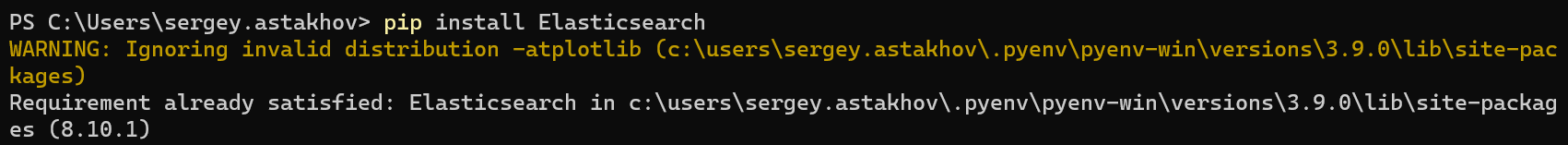


Рисунок 3 — установленный python-пакет

**3. Реализация процесса Data Science на примере диагностики болезней**

Перенесем и проиндексируем данные на сервере ES с помощью программы, код которой приведен в листинге 1.

Листинг 1 — программа для индексации данных

|  |
| --- |
| import json  from elasticsearch import Elasticsearch  # подключение к серверу  client = Elasticsearch([{"host": "127.0.0.1", "scheme": "http", "port": 9200}])  # создание индекса  indexName = "medical"  client.indices.create(index=indexName)  # описание схемы  diseaseMapping = {  "properties": {  "name": {  "type": "text",  "fielddata": True  },  "title": {  "type": "text",  "fielddata": True  },  "fulltext": {  "type": "text",  "fielddata": True  }  }  }  # отправка схемы  client.indices.put\_mapping(index=indexName,  # doc\_type="diseases",  # include\_type\_name="true",  body=diseaseMapping)  # считывание данных из файла  with open('data.json', 'r') as f:  dataStore = json.load(f)  # отправка данных в ES  for data in dataStore:  try:  client.index(  index=data["index"],  # doc\_type=data["doc\_type"],  id=data["id"],  body=data["body"]  )  except Exception as e:  print(e) |

Напишем свой первый запрос в ES (листинг 2, рисунок 4). Будем запрашивать первые 3 подходящих записи, ищя запрашиваемые слова во всех полях документа с различными весами. Результат будет отсортирован по весу (“score”), который зависит от того, в каких полях располагались искомые слова.

Листинг 2 — запрос для поиска по ключевым словам

|  |
| --- |
| indexName = "medical"  docType = "diseases"  searchFrom = 0  searchSize = 3  # 1  searchBody = {  "\_source": ["name"], # возвращать только названия болезней  "query": {  "simple\_query\_string": {  "query": '+fatigue+fever+"joint pain"', # список симптомов  "fields": ["fulltext", "title^5", "name^10"] # веса  }  }  }  result = client.search(index=indexName, body=searchBody, from\_=searchFrom, size=searchSize) |

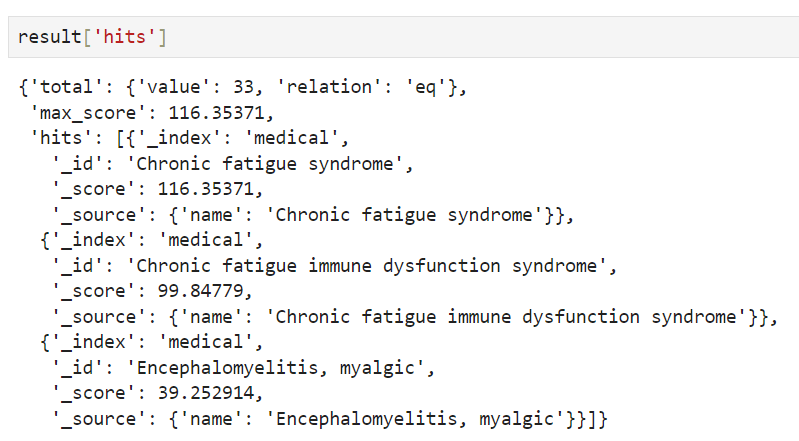


Рисунок 4 — результат выполнения запроса

Первые 3 найденных заболевания: “Синдром хронической усталости”, “Синдром хронической усталости и иммунной дисфункции”, “Энцефаломиелит миалгический”.

Запрос 2 — как изменятся результаты поиска болезней, если в "query" добавить новый симптом rash (сыпь)? Параметры запроса и ответ на него приведены в листинге 3 и на рисунке 5 соответственно.

Листинг 3 — запрос для поиска по ключевым словам

|  |
| --- |
| indexName = "medical"  docType = "diseases"  searchFrom = 0  searchSize = 3  # 3  searchBody = {  "\_source": ["name"], # возвращать только названия болезней  "query": {  "simple\_query\_string": {  "query": '+fatigue+fever+"joint pain"+rash', # список симптомов  "fields": ["fulltext", "title^5", "name^10"] # веса  }  }  }  result = client.search(index=indexName, body=searchBody, from\_=searchFrom, size=searchSize) |

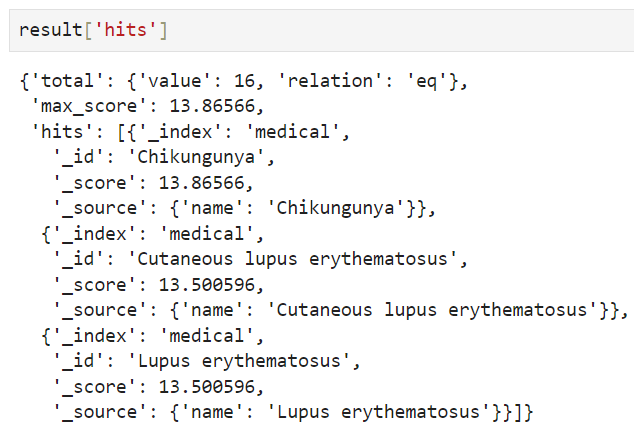


Рисунок 5 — результат выполнения запроса

Как видно, уменьшилось число результатов (value) и изменились их веса (score). Первые 3 соответствующие запросу болезни: “Чикунгунья”, “Кожная красная волчанка” и “Красная волчанка”.

Запрос 3 — Как изменятся результаты поиска болезней, если в "query" добавить ещё один симптом sensitivity to sunlight (чувствительность к солнечному свету)?

Параметры запроса и ответ на него приведены в листинге 4 и на рисунке 6 соответственно.

Листинг 4 — запрос для поиска по ключевым словам

|  |
| --- |
| # 3  searchBody = {  "\_source": ["name"],  "query": {  "simple\_query\_string": {  "query": '+fatigue+fever+"joint pain"+rash+"sensitivity to sunlight"',  "fields": ["fulltext", "title^5", "name^10"]  }  }  }  result = client.search(index=indexName, body=searchBody, from\_=searchFrom, size=searchSize) |

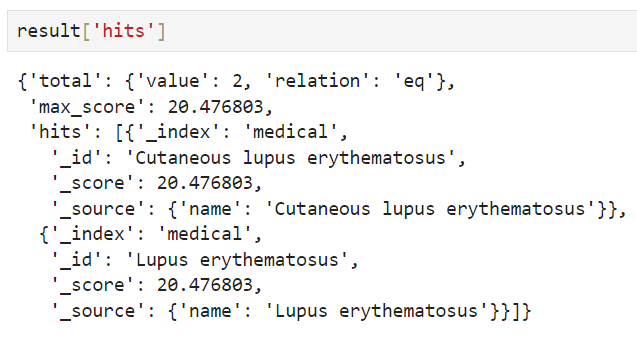


Рисунок 6 — результат выполнения запроса

Как видно, уменьшилось число результатов (value) и изменились их веса (score). Запросу соответствуют всего 2 болезни: “Кожная красная волчанка”, “Красная волчанка”.

Запрос 4 позволяет выявить симптомы болезней, в имени которых встречается слово diabetes (диабет). Параметры запроса и ответ на него приведены в листинге 5 и на рисунках 7-8 соответственно.

Листинг 5 — запрос с агрегацией

|  |
| --- |
| #4  searchBody={  "\_source": ["name"],  "query": {  "query\_string" : {  "query": "name: diabetes"  }  },  "aggregations" : {  "DiseaseKeywords" : {  "significant\_terms" : { "field" : "fulltext", "size" : 10 } # получить частые слова из описания  }  }  }  result=client.search(index=indexName, body=searchBody, from\_ = searchFrom, size=searchSize) |

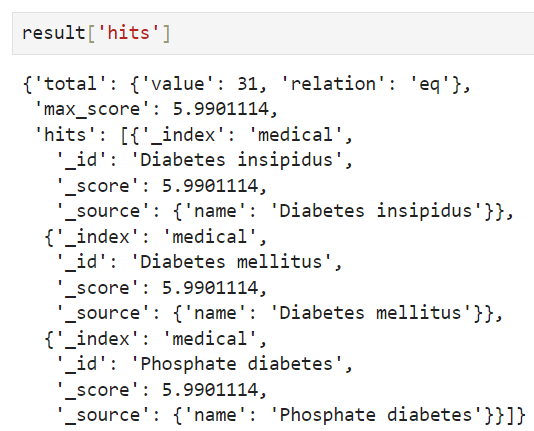


Рисунок 7 — список искомых болезней



Рисунок 8 — список ключевых слов

Какие 5 симптомов диабета являются наиболее важными?

“siphon”, “diabainein”, “bainein”, “ndi”, “passer”.

Сколько документов (названий болезней, связанных с диабетом) было проанализировано?

result['hits'][‘value’] = 31.

Сколько времени выполнялся запрос?

result['took'] = 4884.

Запрос 5 — выявить симптомы болезней, в имени которых встречаются слова diabetes (диабет) и insipidus (несахарный). Параметры запроса и ответ на него приведены в листинге 6 и на рисунках 9-10 соответственно.

Листинг 6 — запрос с агрегацией

|  |
| --- |
| #5  searchBody={  "\_source": ["name"],  "query": {  "query\_string" : {  "query": "name: diabetes+insipidus"  }  },  "aggregations" : {  "DiseaseKeywords" : {  "significant\_terms" : { "field" : "fulltext", "size" : 10 }  }  }  }  result=client.search(index=indexName, body=searchBody, from\_ = searchFrom, size=searchSize) |

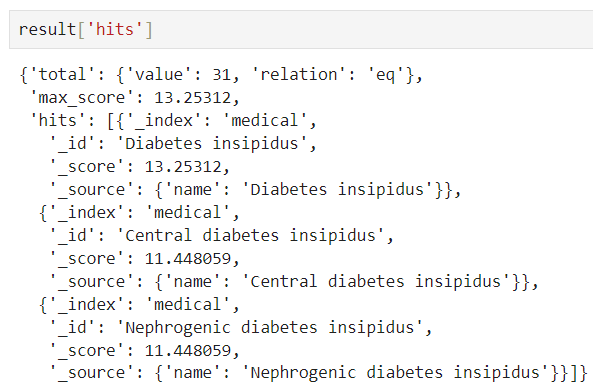


Рисунок 9 — список искомых болезней



Рисунок 10 — список ключевых слов

Запрос 6 — выполните нечёткий поиск (fuzzy) для имени болезни diabetse (в названии болезни случайно поменяли местами две последние буквы). Параметры запроса и ответ на него приведены в листинге 7 и на рисунках 11-12 соответственно.

Листинг 7 — запрос на нечеткий поиск

|  |
| --- |
| #6  searchBody={  "\_source": ["name"],  "query": {  "fuzzy" : {  "name": { "value": "diabetse"}  }  },  "aggregations" : {  "DiseaseKeywords" : {  "significant\_terms" : { "field" : "fulltext", "size" : 10 }  }  }  }  result=client.search(index=indexName, body=searchBody, from\_ = searchFrom, size=searchSize) |

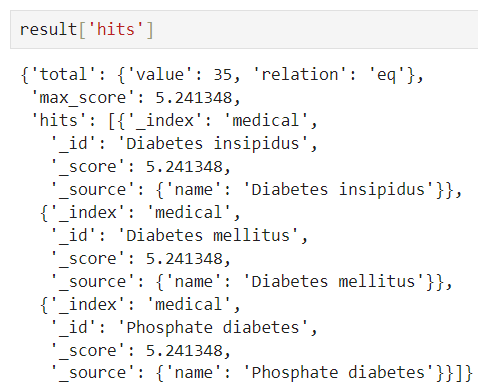


Рисунок 11 — список искомых болезней



Рисунок 12 — список ключевых слов

Как видно, список болезней и ключевых слов не изменились, следовательно, нечеткий поиск работает корректно.

Запрос 7 является примером запроса с фильтрацией (листинг 8, рисунок 13).

Листинг 8 — запрос с фильтрацией

|  |
| --- |
| #7  filterQuery = {  "\_source": ["name", "title", "fulltext"],  "query": {  "constant\_score": {  "filter": {  "term": {  "title": "headache"  }  }  }  }  }  result=client.search(index="medical", body=filterQuery, size=1) |

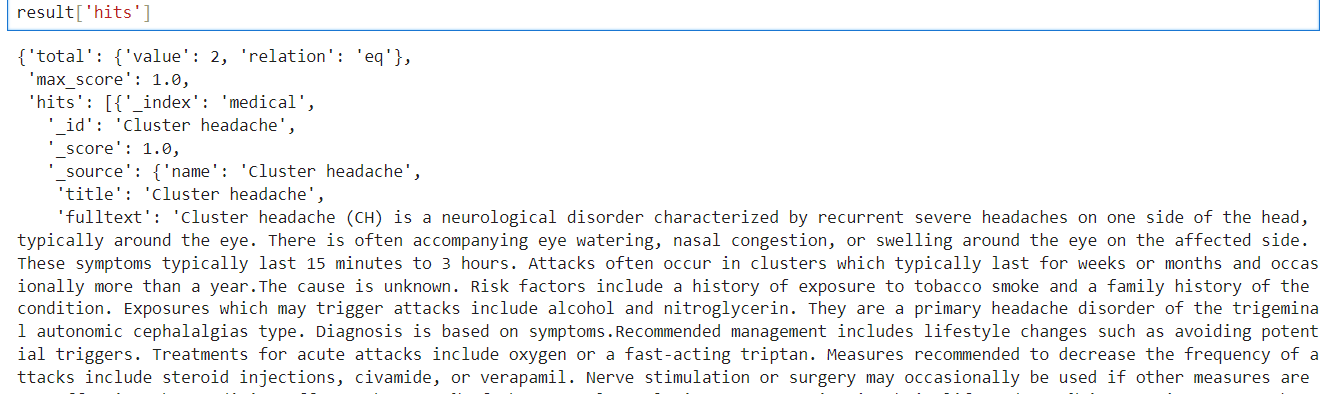


Рисунок 13 — результат запроса

Запрос вернул 2 статьи о головной боли. Начало первой статьи:

“Кластерная головная боль (CH) - это неврологическое расстройство, характеризующееся повторяющимися сильными головными болями с одной стороны головы, обычно вокруг глаза. Часто наблюдается слезотечение, заложенность носа или припухлость вокруг глаза на пораженной стороне. Эти симптомы обычно длятся от 15 минут до 3 часов. Атаки часто происходят группами, которые обычно длятся недели или месяцы, а иногда и больше года.Причина неизвестна...”

Запрос 8 сортирует возвращаемые результаты (листинг 9, рисунок 14).

Листинг 9 — запрос с сортировкой

|  |
| --- |
| sortQuery = {  "\_source": ["name", "title"],  "sort": [{"title": "desc"}]  }  result=client.search(index=indexName,body=sortQuery,from\_=searchFrom, size=searchSize) |

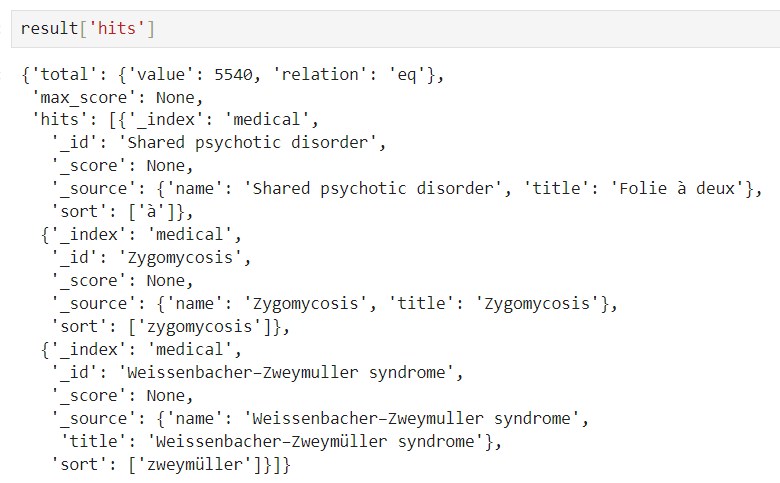


Рисунок 14 — результаты запроса

В результате будут отсортированы все (value = 5540) записи.

Запрос 9 производит извлечение документа по идентификатору (листинг 10, рисунок 15).

Листинг 10 — запрос на извлечение документа по идентификатору

|  |
| --- |
| idQuery = {  "query": {  "terms": {  "\_id": ["Stomatitis", "Alkaptonuria"]  }  }  }  result=client.search(index=indexName, body=idQuery, size=1) |

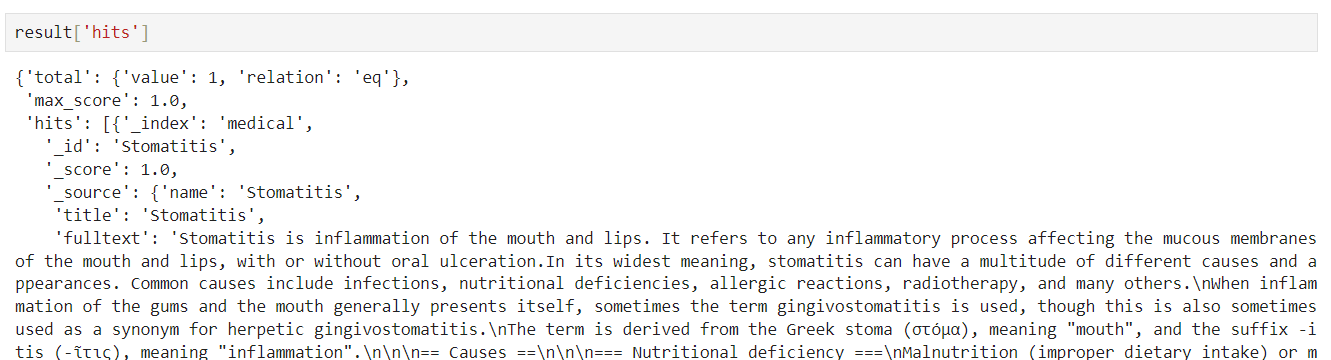


Рисунок 19 — результат запроса

Начало текста статьи: “Стоматит - это воспаление полости рта и губ. Это относится к любому воспалительному процессу, поражающему слизистые оболочки полости рта и губ, с изъязвлением полости рта или без него.В самом широком смысле стоматит может иметь множество различных причин и проявлений. Распространенные причины включают инфекции, дефицит питательных веществ, аллергические реакции, лучевую терапию и многие другие.Когда обычно проявляется воспаление десен и полости рта, иногда используется термин "гингивостоматит", хотя иногда он также используется как синоним герпетического гингивостоматита.\Термин происходит от греческого stoma (στόμα), означающего "рот", и суффикса -itis (-ῖτις), означающего "воспаление"...”

Написать и выполнить запрос 10 bool query, позволяющий найти описания болезней, удовлетворяющих следующим условиям:

1. "must": в поле "title" встречается слово "Chronic"

2. "filter": в тексте документов ("fulltext") встречается слово "Americans"

3. "should": в поле "title" встречается слово "leukemia" или "syndrome"

4. "must\_not": в поле "name" не должно встречаться сочетание "Epidemic encephalomyelitis"

5. Минимальное число низкочастотных терминов (для should), которые должны присутствовать, равно 1 ("minimum\_should\_match": 1).

Текст и результат такого запроса приведены в листинге 10 и на рисунке 16.

Листинг 10 — логический запрос

|  |
| --- |
| boolQuery = {  "\_source": ["name", "title", "fulltext"],  "query": {  "bool": {  "must" : {  "term" : { "title" : "chronic" }  },  "filter": {  "term" : { "fulltext" : "americans" }  },  "should" : [  { "term" : { "fulltext" : "leukemia" } },  { "term" : { "fulltext" : "syndrome" } }  ],  "must\_not" : {  "term" : { "name" : "epidemic+encephalomyelitis" }  },  "minimum\_should\_match": 1  }  }  }  result=client.search(index=indexName, body=boolQuery, size=1) |

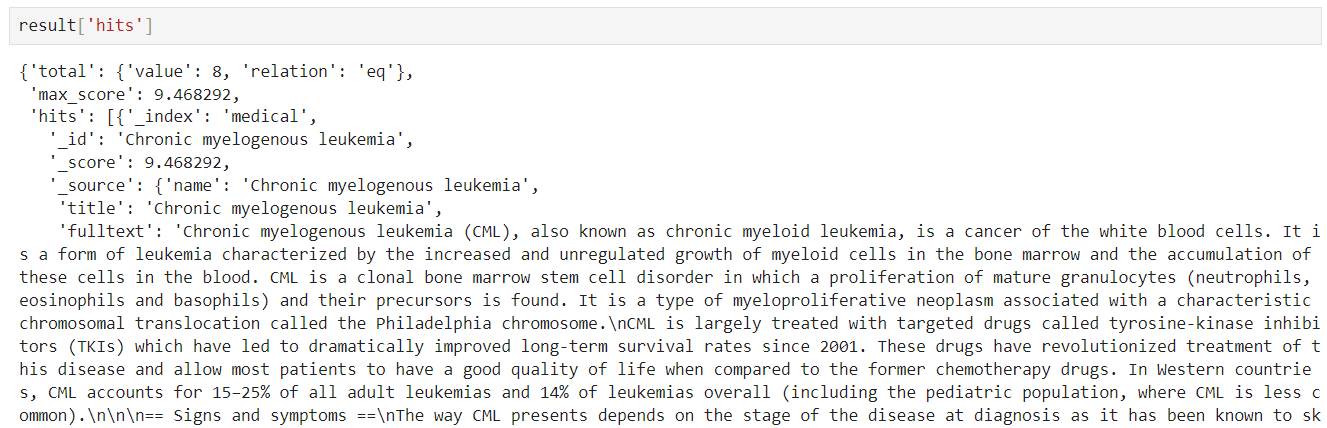


Рисунок 16 — результаты запроса

Запрос вернул 8 заболеваний (value = 8). Начало первой статьи: “Хронический миелолейкоз (ХМЛ), также известный как хронический миелоидный лейкоз, является раком белых кровяных телец. Это форма лейкоза, характеризующаяся усиленным и нерегулируемым ростом миелоидных клеток в костном мозге и накоплением этих клеток в крови. ХМЛ - это клональное заболевание стволовых клеток костного мозга, при котором обнаруживается пролиферация зрелых гранулоцитов (нейтрофилов, эозинофилов и базофилов) и их предшественников...”.

**Вывод:** в ходе лабораторной работы были изучены модель представления данных и способы работы с документной NoSQL базой данных Elasticsearch, получены навыки инсталляции, индексации и поиска в Elasticsearch.