

Федеральное государственное автономное образовательное учреждение высшего образования
«Московский физико-технический институт (национальный исследовательский университет)»
МФТИ



НИЦ «Курчатовский институт»



«Создание алгоритма поиска научных публикаций на основе информации о внешнем цитировании с применением нейросетевых моделей»

Докладчик: Доровских Д.О.

10-я международная конференция
«Распределенные вычисления и Grid-технологии в науке и образовании»
Дубна – 2023 г.

Актуальность

Базы данных научных публикаций в настоящее время насчитывают миллионы статей.

Для увеличения точности поиска важно использовать дополнительные факторы, которые отражают ключевую суть искомой публикации.

- Исследование новых методов поиска по научной литературе: задача семантического поиска научных публикаций на основе информации о внешнем цитировании с использованием нейросетевых моделей по большим базам научных публикаций.
- Источник данных: полнотекстовый архив научных публикаций по биомедицине PubMed Central (PMC) объемом 7,6 миллиона статей (9,1 Тб).

Постановка задачи

- информация о цитировании одних статей другими как независимый фактор
- современные нейросетевые модели на основе алгоритмов трансформеров для поиска по смыслу

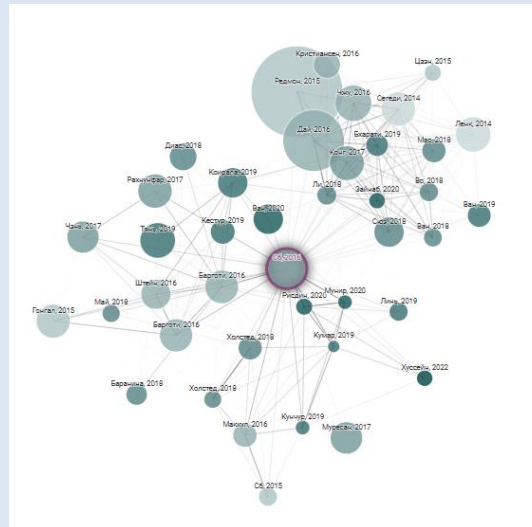
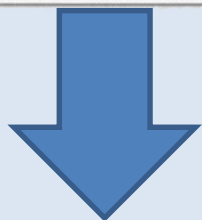


Рис.1 – [\[1\]](#)

[1] Sa I. et al. Deepfruits: A fruit detection system using deep neural networks //sensors. – 2016. – T. 16. – №. 8. – C. 1222.

Постановка задачи

Анализ ДНК методами машинного обучения



Упоминание:

Анализ ДНК и мозговой активности являются типичными примерами, в которых использование методов **машинного обучения** играет важную роль в поиске сложных паттернов

doi:

10.1038/nbt.2450

Название работы:

Дифференциальный анализ регуляции генов при разрешении транскриптов с помощью RNA-seq

Аннотация:

Дифференциальный анализ экспрессии генов и транскриптов с использованием высокопроизводительного секвенирования РНК (RNA-seq) осложнен несколькими источниками изменчивости измерений и создает многочисленные статистические проблемы. Мы представляем Cuffdiff 2, алгоритм, который оценивает экспрессию с разрешением на уровне транскриптов и контролирует вариабельность...

Постановка задачи

Нейросетевые модели для семантического поиска:

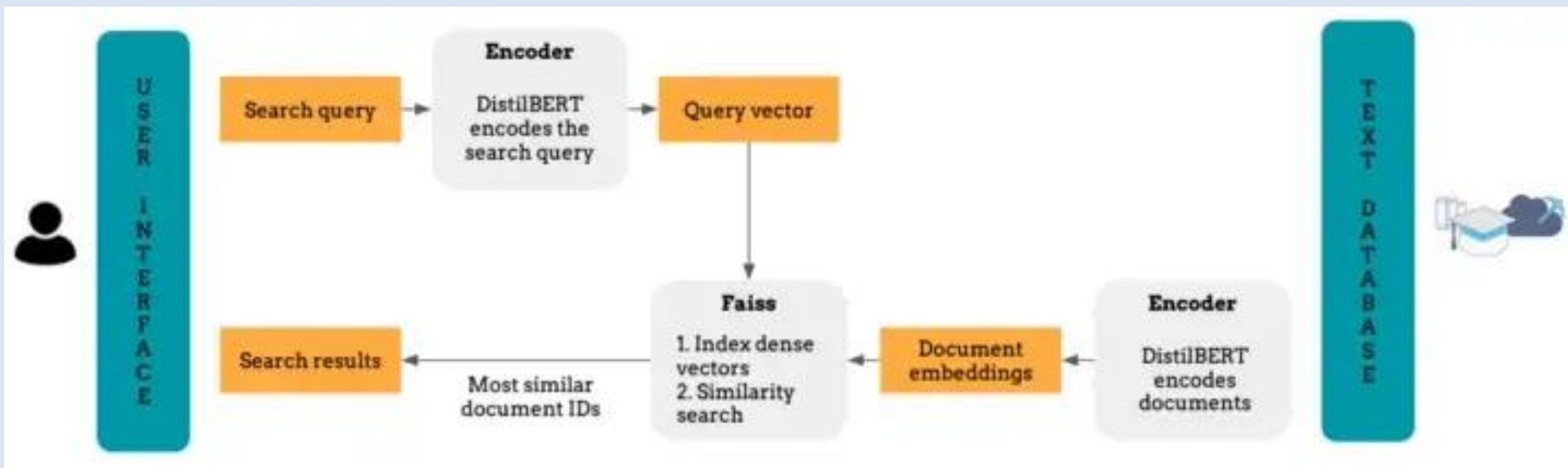


Рис. 2 – [2]

[2] - [How to Build a Semantic Search Engine With Transformers and Faiss | by Kostas Stathouloupoulos | Towards Data Science](#)

Цель и задачи

Цель:

Разработка и тестирование поисковой системы на ресурсах ОВК НИЦ «Курчатовский институт» для интерактивного поиска в архиве научных публикаций.

Задачи:

- разработка методов предобработки/индексирования статей
- создание набора данных с текстовыми упоминаниями публикаций, по которым будет производиться поиск
- дообучение нейросетевой модели BERT для улучшения результатов семантического поиска в БД научных публикаций
- создание макета веб-сервиса для поиска научных публикаций

Результаты

- разработан модуль для обработки научных статей с использованием библиотеки Python lxml: 350000 статей (PMC open access)
- разработаны программные модули для отбора и хранения текстовых упоминаний статей: 550000 записей
- сформированы два набора данных (файлы формата json):



```
{  
  'id':      <id_статьи_на_которую_идет_ссылка>,  
  'id_from': <id_статьи_источника_упоминания>,  
  'quote':   <текст_упоминания>  
}
```

*Запись для
ссылки*

```
{  
  'id':      <id_статьи>,  
  'title':    <название>,  
  'authors':  <авторы>,  
  'abstract': <аннотация>  
}
```

*Запись для
статьи*

Результаты

- разработан программный модуль для поиска публикаций с применением нейросетевой модели Bert
- проведено дообучение нейросетевой модели BertForSequenceClassification для задачи мультиклассовой классификации на наборе из 10000 упоминаний с использованием библиотек transformers, torch, scikit-learn, pandas

Задача оценки качества поиска:

Научная статья

Название:

<название_статьи>

Библиографический список:

[1, 2, ... , n]



Метрика качества поиска:

k/n ,

где k – количество статей из библиографического списка, найденных по результатам поиска

Результаты

1. Запросы подбирались вручную: выполнено **30** запросов

Запрос – набор ключевых слов, выбранных согласно теме и содержанию работы

2. Выбор запросов автоматизирован: выполнено **100** запросов

Запрос - случайным образом выбранное предложение с упоминанием другой публикации в тексте рассматриваемой работы

Оценка средней точности поиска:

	Поиск с применением BERT	Поиск на основе Tf-Idf метрики
запросы подобраны вручную	61%	41%
выбор запросов автоматизирован	58%	39%

Результаты

- поисковый веб-сервис на основе библиотек Flask Python, React Js

Semantic search demo

depression, anxiety, and ap



Id	Title	Authors	Quotes
425958	Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents	Berkman LF, Syme SL	[28] Family and friends play a significant role in recovery from mental illnesses.
1484715	Painful neuropathy—Altered central processing maintained dynamically by peripheral input	Gracely R.H., Lynch S.A., Bennett G.J.	], and chronic pain conditions, arise from altered signaling in both the peripheral and central nervous system [70, 71].
10097003	Peripheral and central hyperexcitability: Differential signs and symptoms in persistent pain	Coderre T.J., Katz J.	], and chronic pain conditions, arise from altered signaling in both the peripheral and central nervous system [70, 71].
10991656	Morphological and functional imaging studies on the diagnosis and progression of Parkinson's disease	Brooks DJ	The severity motor symptoms, especially bradykinesia and rigidity are related to the degree of striatal dopaminergic hypofunction striatal as indicated by an association between FDOPA uptake and "off" stage unified Parkinson disease rating scale (UPDRS) motor score ($r=-0.50$ – -0.68). 7 - 9
		Parisi V., Restuccia R	Recently RNFI thinning in patients with AD [4]. [5]. [6].

Спасибо за внимание!