http://ifgpt.dcl.bas.bg/

Проект

„Инфраструктура за фина настройка на предварително обучени големи езикови модели“

Големият набор от езикови данни IfGPT: управление на метаданните и осигуряване на качеството на данните

Представяне на резултатите от проекта
„Инфраструктура за фина настройка на предварително обучени големи езикови модели“

Ивелина Стоянова
Секция по компютърна лингвистика
Институт за български език, Българска академия на науките

iva@dcl.bas.bg

Основни резултати

- **Голям набор от разнообразни езикови данни** – над 100 млн. документа, обхващащи над 10 млрд. думи.
- **Инфраструктура за осигуряване на качеството на данните** – изчистване, дедупликиране, откриване на пристрастия и др.
- **Разширена схема от метаданни за описание на документите в колекцията** – 15 задължителни характеристики и 8 незадължителни.
- **Представяне на описанието на текстовите единици в графова база данни Neo4J** – метаданните на близо 700 хиляди документа са въведени в базата данни.
- **Уебинтерфейс за достъп до базата данни** – търсене в базата и филтриране по различни характеристики.

Голям набор от данни IfGPT

Общи новини BG

2.1 млн. документа
601 млн. думи

CURLICAT BG

6 хил. документа
22.8 млн. думи

MARCELL BG

25 хил. документа
45 млн. думи

Колекция текстове BG

66 файла
28.9 млн. изречения

Набор от данни IFGPT

над 100 млн. документа
над 10 млрд. думи
над 350 млн. изречения

Лицензи: различни — CC, public domain, ограничени

Периодика (след 1990) BG

4.1 млн. документа
4.4 млрд. думи

Книги BG

22 хил. документа
630 млн. думи

Новини (преди 1990) BG

5.5 млн. документа
270.5 млн. думи

Периодика (преди 199 BG

25 хил. документа
30 млн. думи

електронни корпуси печатни източници

Осигуряване на качеството на данните в IfGPT

Това налага поредица от процедури за осигуряване на последователност и качество чрез създаването на **Инфраструктура за осигуряване на качеството на данните:**

1) **Почистване на текстовете** – премахване на тагове и др.;
2) **Дедупликация** – премахване на повторения и близки повторения (MinHash и Locality Sensitive Hashing, LSH);
3) **Отбелязване на лична информация (PII) и потенциално чувствителна информация** (различни библиотеки и инструменти, например PIISA);
4) **Отбелязване на потенциални пристрастия**.

Осигуряване на качеството на данните в IfGPT

Защо е трудна задачата за оценка на пристрастия?

- **Субективност.** Липсва и единна дефиниция за пристрастие. Възприемането зависи от културния и личния опит на оценителя.
- **Скрити (имлицитни) пристрастия.** Явно маркираните пристрастия (обиди, грубости) се откриват сравнително лесно, за разлика от скритите (стереотипи и културни препратки).
- **Културна и езикова специфика.** Стереотипите са езиково и културно зависими; а повечето изследвания са за английски.
- **Многомерност.** Пристрастието може да се изразява по много параметри (пол, религия, етнос, външен вид, увреждание).
- **Системно разминаване между оценката на пристрастия между хора и големи езикови модели.**

Осигуряване на качеството на данните в IfGPT

Лексикално мотивиран подход

1. Входно изречение

и предварителна обработка

2. Свързване с речник за пристрастия

Сигнални думи

583 единици

Маркиращи думи

3 203 единици

3. Оценка на пристрастието в изречението

въз основа на тежестта на установените двойки (сигнална, маркираща дума)

4. Резултат

отделна оценка на различните видове пристрастие и обща комбинирана оценка

пол | религия | етническа принадлежност | увреждане

Разширена схема от метаданни

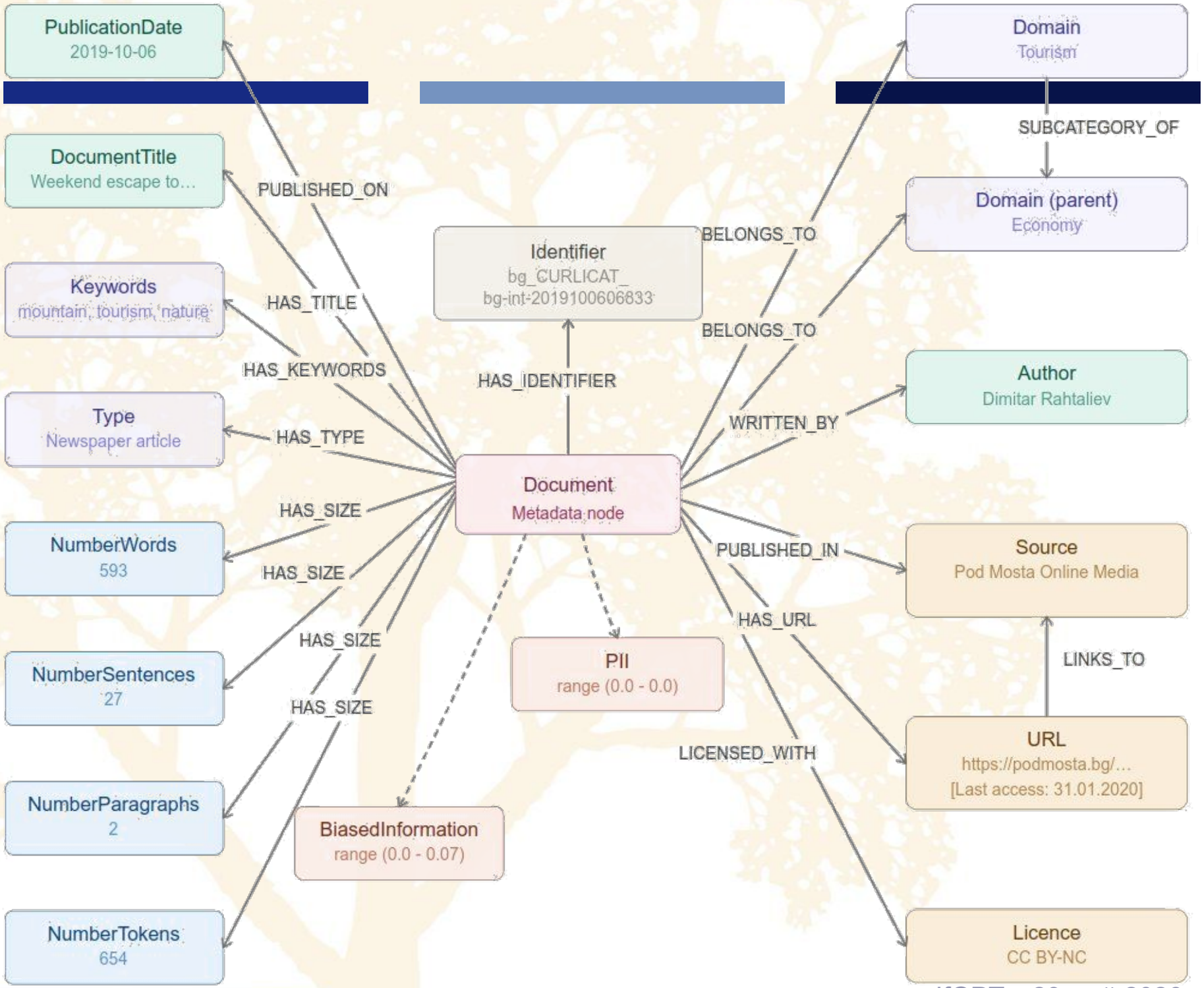
Задължителни полета

Identifier
Licence
PublicationDate
DocumentTitle
Source, Url
Medium
Domain
Keywords
NumberWords
NumberSentences
NumberParagraphs
PersonallyIdentifiableInformation
BiasedInformation

Незадължителни полета

Author
Style
Type
Subdomain
TranslatedDocument
CollectionDate
LicenseLink
TaskCategories

Разширена схема от метаданни

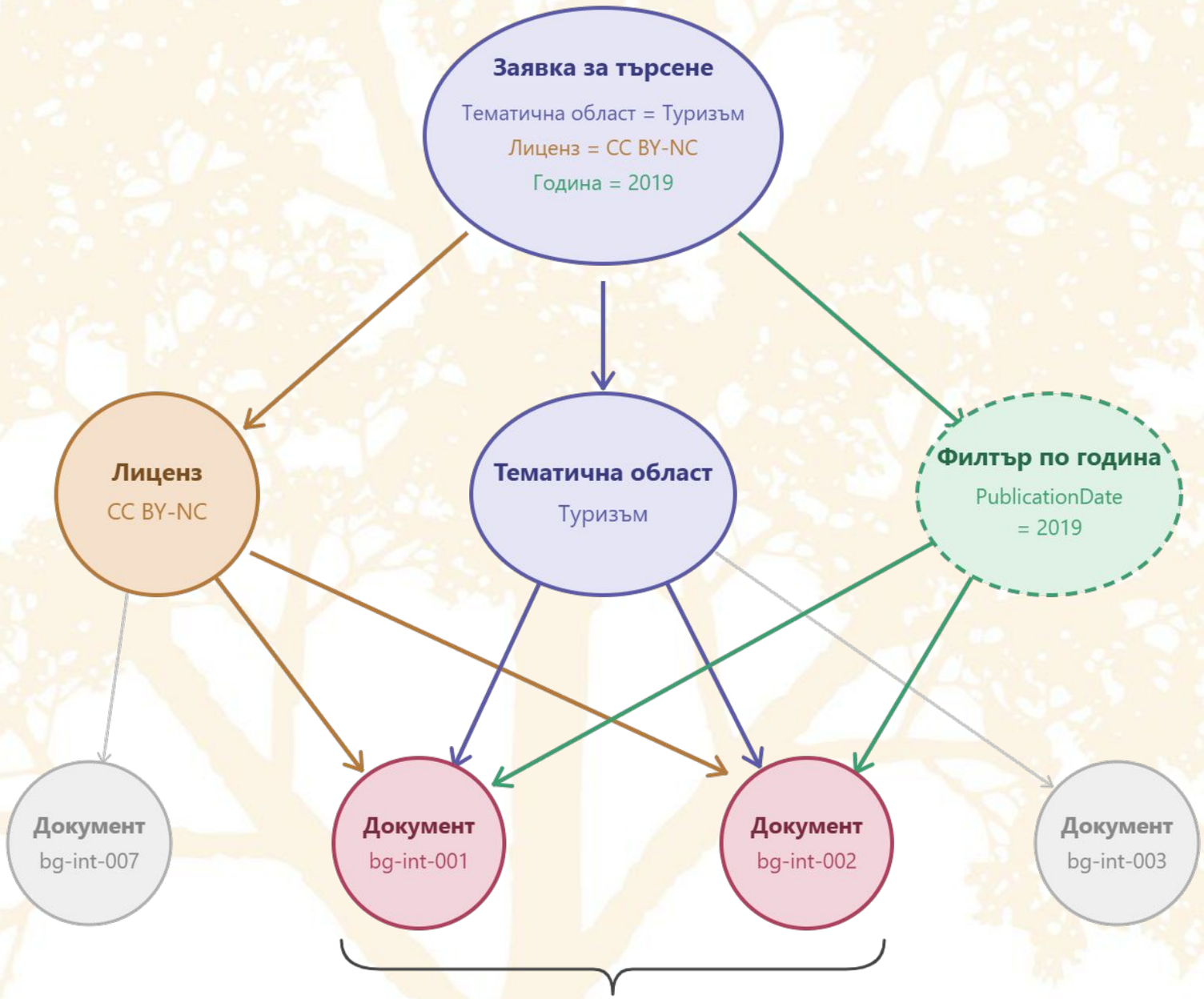


Представяне на метаданните в графова база данни Neo4J

Организацията на отделните категории на метаданните насочва към избора на графовата база данни Neo4J за тяхното съхранение и управление.

- **Единна гъвкава система за хетерогенни данни.**
- **Богата система от релации** за отразяване на връзките между категории в метаданните (например тематични области).
- **Търсене и извличане на информация** по интуитивен и ефективен начин чрез обхождане на граф.

Представяне на метаданните в графова база данни Neo4J



Уебинтерфейс за достъп до метаданните

https://ifgpt.dcl.bas.bg/ifgpt-dataset/

Версия: 1.0 (Open Beta) | история на версиите

Търсене в 689,645 документа

Тип лиценз

Избери всички | Изчисти всички

- ☐ Свободен
- ☐ Ограничен

Лиценз

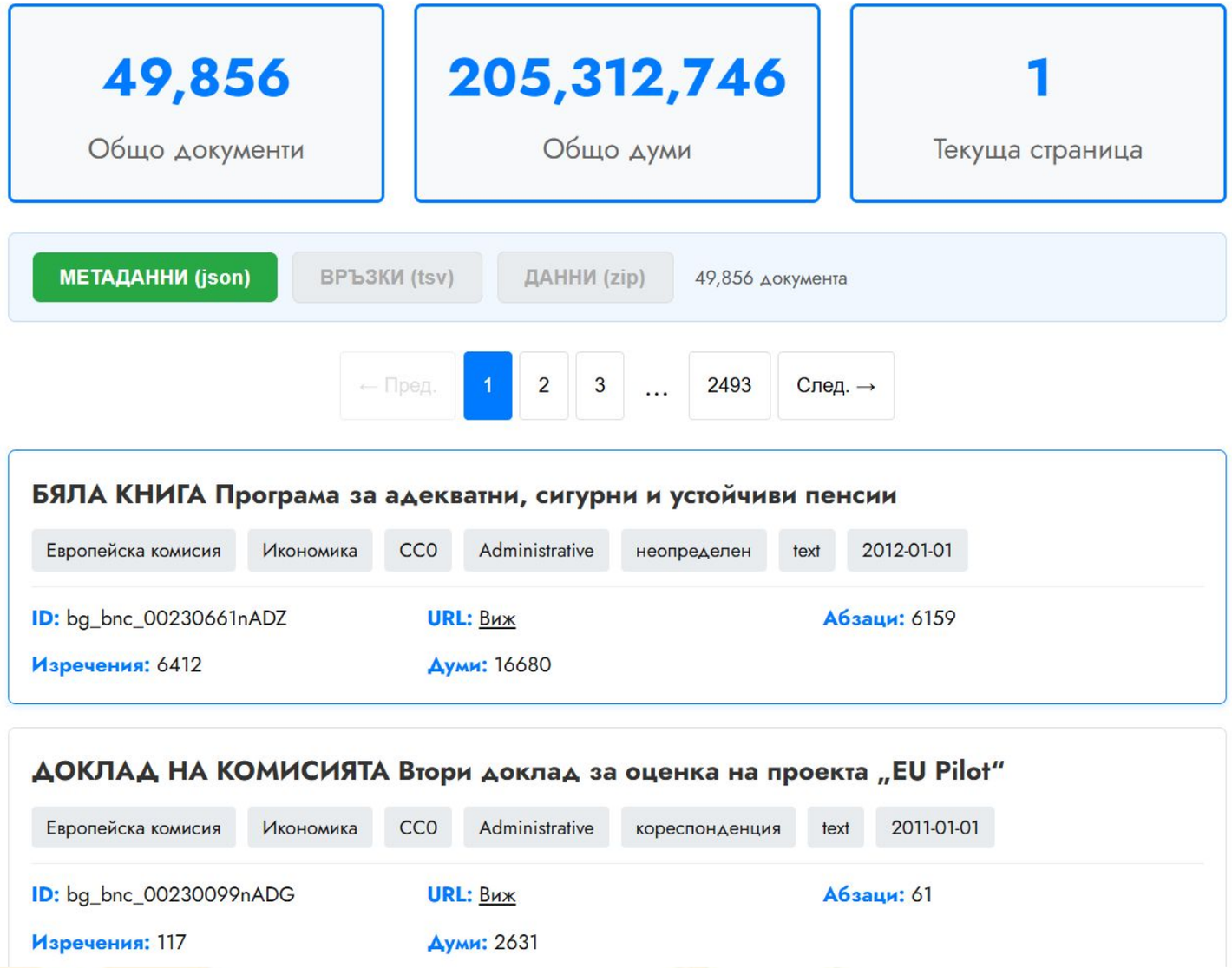
Избери всички | Изчисти всички

- ☐ BTA Licence
- ☐ CC BY 2.5
- ☐ CC BY 4.0
- ☐ CC BY-NC
- ☐ CC BY-NC 4.0
- ☐ CC BY-NC-SA 2.0
- ☐ CC BY-NC-SA 2.5
- ☐ CC BY-NC-SA 3.0
- ☐ CC BY-NC-SA 4.0

Уебинтерфейс за достъп до метаданните

Търсене и филтриране по:

- Тип лиценз и конкретен лиценз;
- Тематични области;
- Времеви период;
- Ключови думи.



IfGPT: Бъдещи насоки

- Увеличаване на големия набор от данни IfGPT.
- Усъвършенстване на процедурите за проверка и оценка на качеството на данните.
- Поддържане и развиване на графовата база данни и интерфейса за достъп.
- Разширяване на описанието с метаданни, релациите между отделните категории, йерархични и други отношения между тях.

Благодарности

- Работата е част от проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“ по договор номер ПВУ – 55 от 12.12.2024 /BG-RRP-2.017-0030-C01/.

Финансирано от Европейския съюз
NextGenerationEU

Национален план за възстановяване и устойчивост
НА РЕПУБЛИКА БЪЛГАРИЯ

- https://ifgpt.dcl.bas.bg