Софийски университет “Св. Климент Охридски”

Факултет по математика и информатика

**Курсов проект**

По "Семантичен уеб”

Трансформация на текст в граф

Изготвили:

Стоян Николов, Технологии за големи данни

Николай Шиваров, Технологии за големи данни

Нели Узунова, Изкуствен интелект

Съдържание

[Въведение](#_faebrgoeoo7m)

[Основни функционалности](#_bsk2qv25oxco)

[1. Въвеждане на текст](#_oz9ja6js5aty)

[2. Избор на схема](#_b5y7h66x6oqz)

[3. Валидация на Turtle формат](#_i0lpsgbctz2d)

4. Премахване на невалидни класове и предикати при изпозването на schema.org

[5. Генериране на граф от знания](#_cxnfmi1jjlg)

[Примери за употреба](#_wojz4t4sfxvd)

[Заключение](#_2cwkl9nyspgw)

## Въведение

Целта на проекта е да се разработи софтуерно приложение, което автоматично трансформира неструктуриран текст в структуриран граф от знания, съобразно зададени схеми като schema.org или FHIR. Идеята е да структурира и визуализира текстовата информация във формат, който улеснява разбирането на връзките и понятията в текста.

## Основни функционалности

#### 1. Въвеждане на текст и извличане на връзките между обектите в него и извеждането им в turtle формат

Потребителите въвеждат текстова информация в системата. Този текст се обработва автоматично, за да се извлекат ключови понятия и връзки. Това става с помощта на голям езиков модел. В случая сме използвали llama3-70b чрез LlamaAPI. За генериране на отговора сме използвали следните инструкции:

* For the given text provide all concepts and relations between them in turtle format. Use Rdfs schema, XML schema, schema.org. In addition for concepts use example.org.
* For the given text provide all concepts and relations between them in turtle format. Use Rdfs schema, XML schema, FHIR. In addition for concepts use example.org and mappings to icd-10.

Като след това се добавя текста, който потребителя е въвел. Големият езиков модел обработва текста и връща отговор в turtle формат, който подлежи на по нататъшна обработка.

#### 2. Избор на схема

При въвеждането на текст потребителите могат да изберат между различни схеми чрез радио бутоните в потребителския интерфейс. Схемите, които ние ползваме са schema.org за общи данни и FHIR за медицински данни. Тези схеми определят как данните ще бъдат структурирани и представени.

#### 3. Валидация на Turtle формат

Често големият езиков модел при генерирането на turtle формата за дадения текст допуска грешки в синтаксиса му. За това имаме няколко функции за справяне с грешки, които сме срещали по време на работата ни. Например за добавяне на префикси, които са използвани, но не са налични в началото на файла или на неправилно сложени шпации или препинателни знаци.

#### 4. Премахване на невалидни класове и предикати при изпозването на schema.org

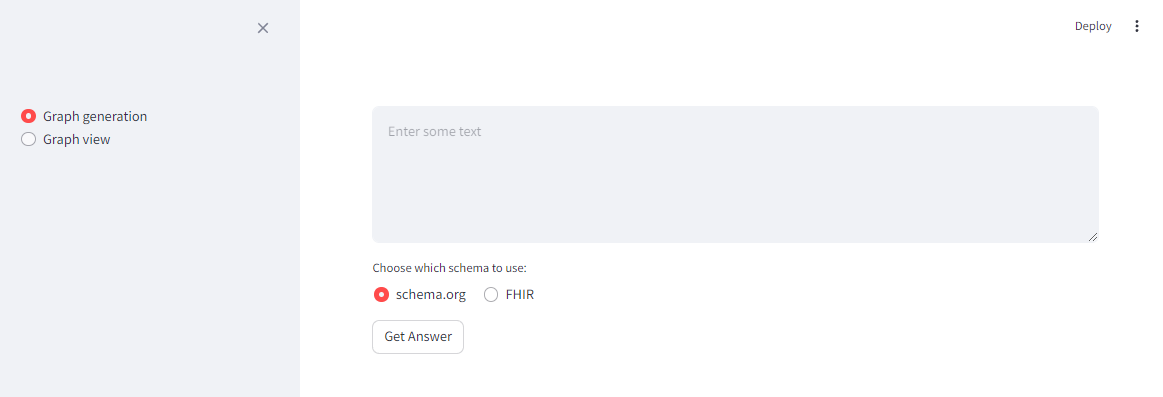
Също така при генерирането на turtle файла, големият езиков модел си измисля обекти, които не са от тези на schema.org. Справяме се с това, като сме свалили оригиналната схема и от нея сме извлекли всички класове и предикати, които съдържа. След това с определена близост до тях съпоставяме тези които езиковия модел ни е генерирал и ги коригираме, ако има грешка в изписването или ги отхвърляме, ако няма такъв обект в схемата.

#### 5. Генериране на граф от знания

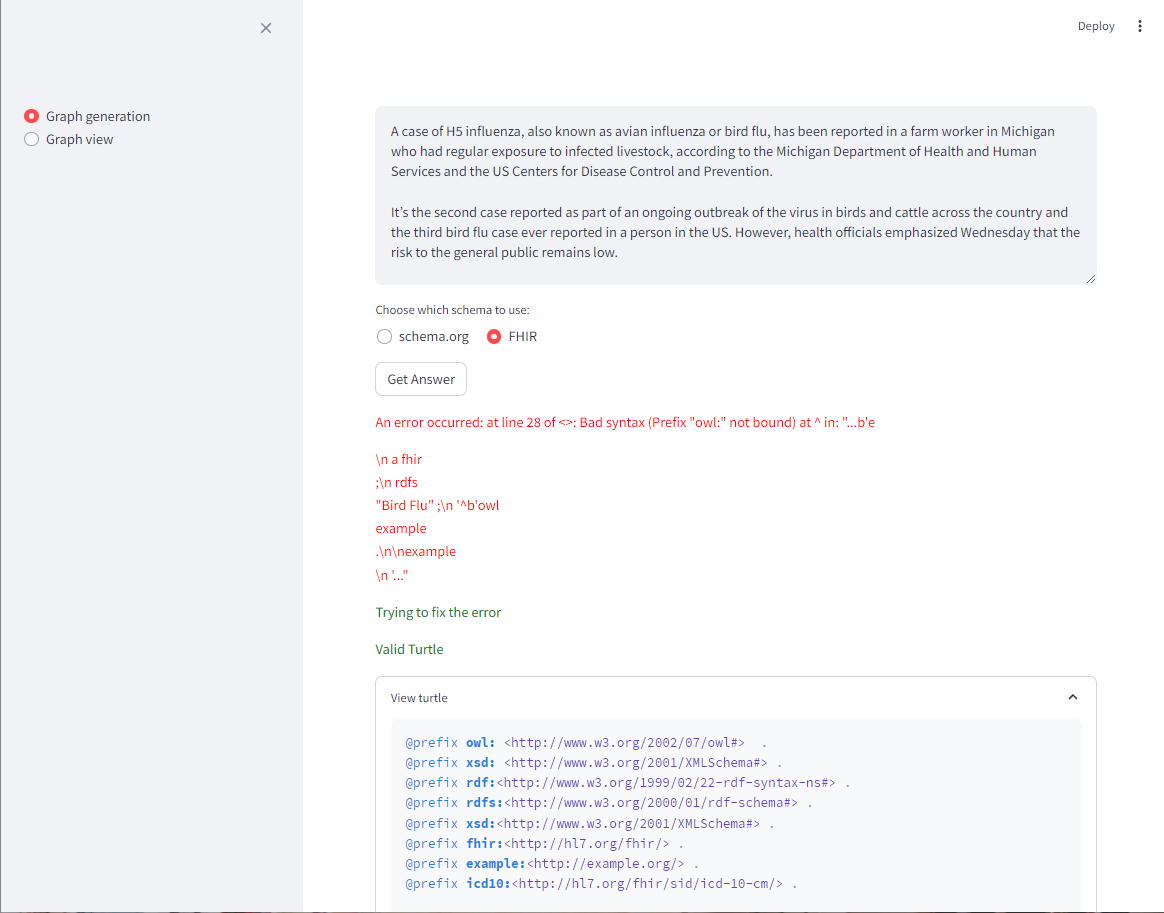
След като текстът е валидиран системата автоматично генерира граф от знания, който визуализира връзките между различните понятия и термини в текста.

## Примери за употреба

Визията на потребителския интерфейс изглежда така:

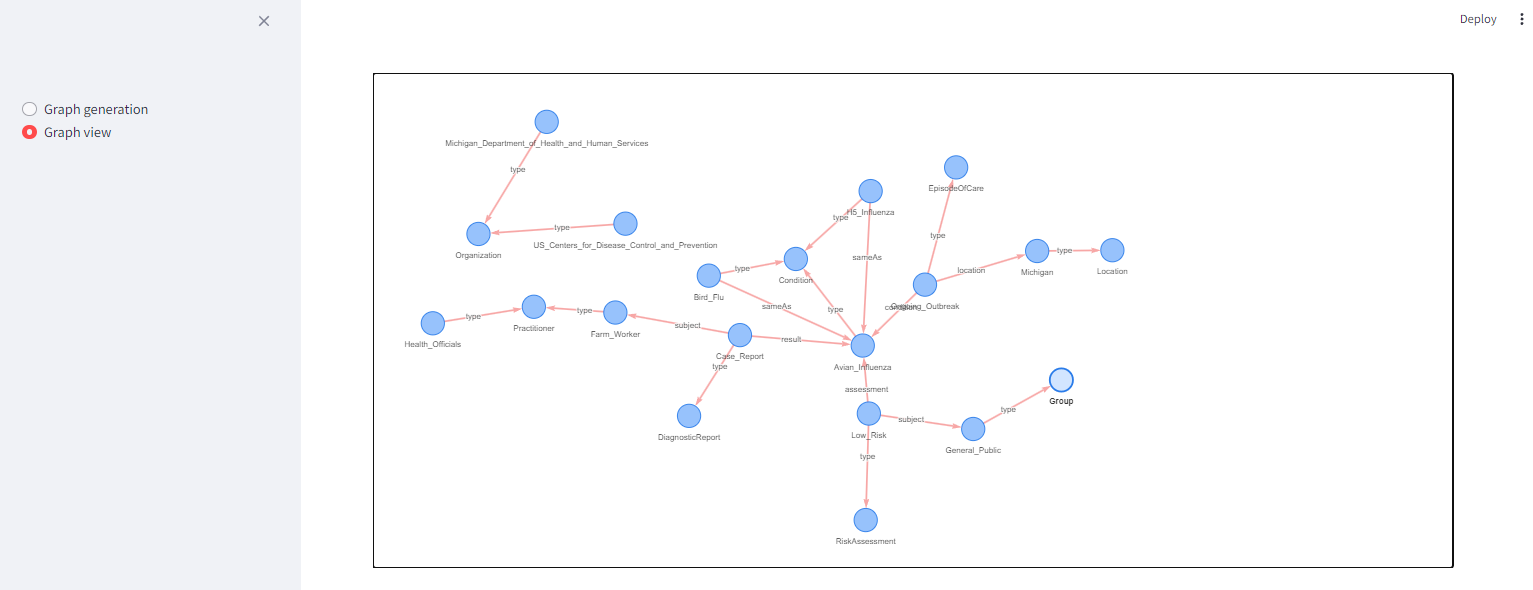


Потребителя може да въведе желания текст, и да избире схема между schema.org или FHIR и да види неговия текст превърнат във валиден turtle формат:





На другата страница може да се види интерактивния граф, отговарящ на желания текст.



## Заключение

Проектът предоставя функционално приложение за автоматична трансформация на текст в граф от знания, използвайки зададена схема. Това улеснява структурираното представяне на текстови данни и подобрява анализирането и използването им в различни приложения.