Специфікація ПЗ

1 ВСТУП

1.1 Огляд ПЗ

Цей документ описує специфікацію програмного забезпечення для дослідження оптимальних параметрів RAG системи та новітнього методу узгодження моделі з базою даних користувача під час візуальних питань-відповідей.

1.2 Мета

Основною метою створення цього документу є опис загальних вимог до програмного забезпечення, включаючи детальний опис функціоналу програмної системи.

1.3 Межі

Розробка вищезазначеної програмної системи включає в себе наявність наступного функціоналу:

– чат інтерфейс для взаємодії з моделями;

– отримання передбачень у якості відповіді на запитання;

– відображення результату передбачення у інтуїтивно зрозумілому для користувача вигляді;

– можливість для завантаження даних в базу даних.

1.4 Посилання

Документ містить посилання на багато джерел дослідницьких результатів, таких як «arXiv.org». Також наведено посилання на паперові видання та інтернет-джерела. Також, документ посилається на роботи викладачів кафедри ПІ та загалом університету ХНУРЕ.

1.5 Означення та абревіатури

ПЗ – програмне забезпечення

ПС – програмна система

2 ЗАГАЛЬНИЙ ОПИС

2.1 Перспективи ПЗ

Розроблена програмна система має великий потенціал для розвитку, вона може бути розглянута з точки зори ще одного кроку до кращого узгодження великих мовних моделей з специфічними даними користувачів. Також , ПЗ може бути використане як частина функціоналу по підбору гіперпараметрів в RAG системах.

2.2 Функції ПЗ

Основними функціями програмного забезпечення є:

– чат інтерфейс для взаємодії з моделями;

– отримання передбачень у якості відповіді на запитання;

– відображення результату передбачення у інтуїтивно зрозумілому для користувача вигляді;

– можливість для завантаження даних в базу даних.

2.3 Характеристики користувачів

Основними користувачами даного програмного забезпечення є дослідники, інженери, які знаходяться в пошуках інноваційних або вже пророблених рішень для узгодження моделей з специфічними даними.

2.4 Загальні обмеження

Програмна система має наступні обмеження:

– розмір датасетів для перевірки підходів є достатньо невеликим;

– час відповіді від БД є досить високим через використання безкоштовного екземпляру;

– інтегровано лише моделі від Google та OpenAI;

– якість відповідей сильно залежить від даних та моделі;

– наявність доступу до мережі інтернет (хоч і не завжди).

2.5 Припущення й залежності

Розроблена ПС має наступні припущення та залежності.

Припущення:

– інженери та дослідники зацікавлені у використанні нових технологій;

– потреба в інтеграції специфічних даних користувачів стає все більшою.

Залежності:

– від апаратного та програмного забезпечення;

– від стабільності доступу до мережі Інтернет.

3 КОНКРЕТНІ ВИМОГИ

3.1 Вимоги до зовнішніх інтерфейсів

3.1.1 Апаратний інтерфейс

Основною обчислюваної одиницею є сервер або комп’ютер користувача, він безпосередньо приймає запити від користувача, оброблює їх та відправляє назад відповіді. Через те, що дана програмна система містить нейронну мережу, то сервер повинен обов’язково мати відеокарту для ефективного виконання поставлених перед ним задач.

Також найкращим вибором буде встановити на сервер операційну систему Linux, наприклад Ubuntu 22.04 LTS, яка відповідає усім вище описаним потребам та є досить частим рішенням в подібних системах.

3.1.2 Програмний інтерфейс

Серверна частина програмної системи повинна мати встановлену операційну систему Ubuntu 22.04 LTS, на якій має бути встановлений python3.11 та необхідні бібліотеки для використання моделей машинного навчання (PyTorch, Transformers).

Програмне забезпечення може бути запущене з використанням скриптів або ноутбуків.

3.1.3 Комунікаційний протокол

Основним протоколом взаємодії між клієнтом та серверами з моделями є протокол HTTP.

3.1.4 Обмеження пам’яті

Сервер для повноцінного та ефективного виконання поставлених операцій має містити від 64Gb фізичної пам’яті та від 24Gb відео-пам’яті.

3.1.5 Операції

Розроблювана програмна система повинна виконувати наступні операції:

– взаємодіяти з користувачем через будь-який інтерфейс;

– отримання передбачень у якості відповіді на запитання;

– відображення результату передбачення у інтуїтивно зрозумілому для користувача вигляді;

– можливість для завантаження даних в базу даних.

3.1.6 Функції ПЗ

Першою компонентою програмної системи є підбір гіперпараметрів для RAG систем. Виходячи з цього, вимоги до розроблюваної частини виглядають наступним чином:

– як результат передбачення, модель має в зрозумілій формі відповідати на запитання користувача;

– модель повинна розуміти синтаксис англійської мови (в ідеалі ще багатьох інших);

– у якості фреймворку для оцінки якості моделей буде використано DeepEval;

– має бути написана на мові програмування Python;

– база даних має використовуватись у випадках коли це потрібно для відповіді;

– дані для перевірки мають бути синтетичними (які модель раніше точно не бачила).

Другою компонентою програмної системи є інноваційний метод узгодження великої лінгвістичної моделі з обмеженнями користувача. Вимоги до розроблюваної частини виглядають наступним чином:

– точність відповідей на специфічному датасеті має бути більше за 80%;

– у якості фреймворку для оцінки якості моделей буде використано TruLens;

– задачу має виконувати мультиагентна система з використанням фреймворку smolagents.

3.1.7 Припущення й залежності

Розроблювана ПС має наступні припущення та залежності.

Припущення:

– інженери та дослідники зацікавлені у використанні нових технологій;

– потреба в інтеграції специфічних даних користувачів стає все більшою.

Залежності:

– від апаратного та програмного забезпечення;

– від стабільності доступу до мережі Інтернет.

3.2 Властивості ПЗ

Розроблювана програмна система повинна мати ряд властивостей для успішного розповсюдження:

– масштабованість, так як основні розрахунки виконуються на серверній частині, то вона має вміти адаптуватися під можливі запити користувачів;

– доступність, відкритість до комерційного використання методів та алгоритмів;

– якість, основна функція даної програмної системи це передбачення відповідей для користувачів, тому ці відповіді мають бути досить гарними для того щоб для користувача мало сенс користуватися цим ПЗ;

– гнучкість, дана програмна система повинна бути написана зрозумілим та доступним для розробників кодом, який, у майбутньому, було б зручно використовувати для свої потреб та розробок у сфері машинного навчання.

3.3 Атрибути програмної системи

3.3.1 Надійність

Наразі програмна система є повністю локальною. База даних має підтримувати переривання операції та швидше їх відновлення, у разі виникнення проблем з мережею.

3.3.2 Доступність

Програмна система повинна швидко реагувати на дії користувача. Допустимим є час очікування рівний 10 секундам.

3.3.3 Безпека

Дані користувачів повинні бути захищеними при передачі та не мають використовуватися в будь-яких інших потребах.

3.3.4 Супроводжуваність

Задача розробника це підтримувати та супроводжувати цілісність системи, та відповідати на можливі запитання користувачів через GitHub.

3.3.5 Переносимість

Розроблювана програмна система має бути адаптована до великого спектру операційних систем. Користувач повинен мати змогу використати наше ПЗ з Linux, Windows, MacOS, Android або iOS, або будь-яких інших операційних систем. Програмне забезпечення також повинно бути доступним з багатьох браузерів, таких як: Google Chrome, Safari та Opera.

3.3.6 Продуктивність

Час відповіді системи на запит користувача не повинен перевищувати 10 секунд.