АЛГОРИТМ АНАЛІЗУ ПРЕДМЕТНОГО ТЕКСТУ ТА СИНТЕЗУ РЕЧЕНЬ НА УКРАЇНСЬКІЙ МОВІ

Мазурок І. Є., Шляхов Д. В., Колбасюк В. О.
Одеський національний університет імені Мечникова

У нашій роботі розглядається задача аналізу предметного тексту та синтезу речень на українській мові з урахуванням зустрічаємості переходів слів у тексті.

Ключові слова: аналіз предметного тексту, синтез речень, алгоритм аналізу текстових данних, частота переходів, генерація речень та текстів, регулярні вирази, контейнер unordered_map, машинне навчання.

Наша робота базується на статистичному аналізі текстових даних та генерації нових речень або текстів на основі цих даних. Ми розробили програму, яка приймає текстовий ввід та аналізує його, зберігаючи статистику вживання слів та переходів між ними. Для цього ми використовували контейнери std::unordered_map, які дозволяють швидкий доступ до даних та ефективний пошук. Також під час виконання нашого проекту ми також використовували регулярні вирази для обробки тексту. Регулярні вирази представляють собою спеціальну мову, яка використовується для пошуку та маніпулювання текстовими рядками.

Програма була написана мовою програмування C++, яка є ефективною та швидкою мовою програмування, що дозволило нам швидко та точно обробляти великі обсяги даних.

Також ми використовували бібліотеку regex у мові програмування C++, яка надає зручні інструменти для роботи з регулярними виразами. Зокрема, ми використовували функцію std::regex_replace для видалення деяких символів зі слів перед їх аналізом та генерацією тексту.

Регулярні вирази дозволяють нам більш точно та ефективно обробляти текст, виділяти з нього потрібні елементи та проводити аналіз даних. Використання

регулярних виразів у нашому проекті значно спростило обробку вхідних даних та підвищило точність генерації тексту.

Для синтезу нового тексту ми використовували алгоритм зваженого випадкового вибору, який враховує частоту вживання слів та їх переходів.

На основі наших досліджень ми зробили висновок, що використання статистичного аналізу та алгоритмів зваженого випадкового вибору є ефективним підходом до генерації нового тексту на основі вхідних даних. Наша програма може бути використана в різних галузях.

Наприклад:

- Duolingo. Наша програма здатна проводити аналіз тексту та створювати невірні варіанти відповіді разом з правильними, які мають подібний контекст. Це може створити виклик для користувача при виборі правильної відповіді, збільшивши рівень складності завдання.

Цей алгоритм можна використовувати у будь-яких подібних додатках, де треба створювати невірні варіанти віповіді за темою питання.

- Генерація діалогів. Цей код можна використовувати для генерації випадкових текстів, що мають подібний контекст до англомовних переговорів пілотів з аеропортом. З такими згенерованими текстами можна навчати пілотів, як ефективно і точно спілкуватися в подібних ситуаціях, або ж використовувати їх для тестування знань пілотів відповідно до міжнародних стандартів. Такий підхід може бути корисним для підготовки пілотів до реальних переговорів, а також для покращення їх навичок взаємодії з контролерами в повітряному просторі.

- Генерація диктантів. Наша програму можно навчинити на типових простих реченнях та інших текстів і генерувати диктанти для разного віку, рівня та лексики.

- Створення моделі машинного навчання. Наша програма може бути використана для покращення розуміння запитів людей та контексту, у якому вони вживаються. Аналіз текстових даних та визначення частоти вживання слів та їх контексту може допомогти створити модель машинного навчання, яка може бути використана для поліпшення розуміння запитів людей, які містять подібні слова або фрази. Наприклад, якщо програма машинного навчання побачила багато запитів про погоду, вона може створити модель, яка розуміє, що питання "Яка погода сьогодні?" пов'язане зі словами "погода" та "сьогодні". Це може допомогти машинному інтелекту краще розуміти контекст запиту та відповідати на нього більш точно.

Приклад генерації диктантів:

Аж ось він навіки заснув у сяєві? Лилик сидів тихо в своєму кутику, ні серце не бачив світла, душею тільки свиснув йому летіти на вогонь? Сидів він би тее світло спалило його, так і почав кружляти понад лампою, що раз то ж бачить смерть у самий поломінь. отеє ж його згуба! і полинув за столом велике товариство. Воно го рить, миготить, міниться, там самотнім, мав-таки сусіда сусід той був ледве примітний тоненький, як і перш е. отеє ж бачить смерть у темнім льоху? Летіти шукати того блискучого проміння і сили. Він затріпотів крил ьцями й не мав відваги і се було в льоху, падав блідесенький промінь, та до згубливого світла. Хіба лилик, так швидко, скільки сили було в темному вогкому льоху тее світло й полетів та все шле свої темні крильця. Метелик на світло, але він, що не мав відваги і розібрати, що було мало! дурному дурна й смерть! Метелик л етить все ближче, ближче до того ще темнішого, щоб міг метелик полетів та порада? Світло, світло! Хто вели в йому було, метелик так ні, таки лізе! Метелик на світ той був неговіркий, понурий собі, та порада? отеє ж бачить смерть у сяєві? Хтось із товариства хотів його прогнати з того ще з того ще з того світла не бачи ти. Хто ж розумніша була користь для метелика: лилик тільки свиснув йому й меншими й хотів він би у великі й кімнаті там життя стратить? Лампа спалахнула, а порадитися ні за бочкою нахилявся, набираючи капусти, і се було в льоху, падав блідесенький промінь, та з ким, бо часом з льоху. Метелик летить все ближче, ближче до льоху за бочку й смерть! На столі була користь для метелика: лилик тільки кутка ще далі за столом велик е товариство. того ні думка, ні серце не був лилик був лилик був неговіркий, понурий собі, та блідий, мов погляд недужої дитинки. У темному вогкому льоху тее світло й заснув нічого й меншими й смерть! того ні за ним лилик був ледве примітний тоненький, як і почав кружляти понад лампою, щораз то ж бачить смерть у темр яві самотній, та не міг сидіти там життя!

Література

1. GitHub посилання на проект– 2023– Посилання: https://github.com/shliakhovdan/text-analysis-and-sentence-synthesis-ukr