ВИКОРИСТАННЯ МАРКОВСЬКИХ МОДЕЛЕЙ ДЛЯ ІНФОРМАЦІЙНОГО ПОШУКУ У ТЕКСТАХ

Сивопляс І.М., Черних О.П., Шеін О.М.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

У XXI столітті інформація займає найбільш важливе місце серед усіх інших ресурсів. Завдяки Інтернету інформації стало настільки багато, що неможливо обробити її усю і потрібно виділити важливі частини. Для цього необхідно використовувати інформаційний пошук. Інформаційний пошук − це пошук неструктурованої інформації у документах, текстах, реляційних базах даних та гіпертекстових базах даних, таких як Інтернет та локальні мережі.

Для швидкого інформаційного пошуку застосовують складні системи пошуку та «вирізання» потрібної інформації з великого об’єму даних. Однією з таких систем є бібліотека StanfordNLP, що вирішує багато проблем у сфері обробки природної мови, а також проблему інформаційного пошуку у глобальній мережі, якій приділяється увага у даній роботі.

Спочатку потрібно обрати мову програмування, у нашому випадку це C#, а потім скачати і встановити останню версію бібліотеки StanfordNLP. Далі потрібно обрати із бібліотеки потрібні класи для використання. Після успішного встановлення та обрання потрібних частин бібліотеки можна приступати до аналізу задачі і налаштування моделей під неї. Для деяких задач є вже налаштовані моделі, які можна знайти у мережі Інтернет, але для більшості задач все ж таки доводиться виконувати тренування заново. Для цього обирається порівняно малий корпус ~10000 текстів припустимо із вже виділеними потрібними ключовими словами та взаємозв’язками між ними. Ці тексти діляться на тренувальний сет (більшість даних) та контрольний сет (малий залишок) і проводиться тренування моделі на тренувальному сеті. Потім виконується перевірка правильності тренування моделі за допомогою контрольного сету даних. Після правильного навчання отримуємо модель, що може шукати інформацію за певною темою у дуже великих об’ємах текстів, обробити які людина самотужки не в змозі.

Для розробника платформа StanfordNLP є корисною, так як надає багато можливостей для обробки і пошуку інформації, та досить простою у застосуванні. Оскільки її також розробляють видатні науковці у сфері обробки природної мови, то методи, що застосовуються у платформі, постійно оновлюються і вдосконалюються.