

УДК 004.4:004.6

Чернов Д.С.¹, Ільяшенко М.Б.²

¹ студ. гр. КНТ-622м НУ «Запорізька політехніка»

² канд.техн.наук, доц. НУ «Запорізька політехніка»

АЛГОРИТМ СЕМАНТИЧНОГО ПОШУКУ

Однією з головних проблем у використанні цінності текстових даних є неструктурованість природної мови. Цей формат даних має сенс для нас, людей, але його дуже складно обробляти автоматизованим способом. Однак, незважаючи на цю проблему, приблизно 80–90% даних на підприємстві є текстом. Саме тут з'являється інформаційний пошук, і зокрема семантичний пошук, який є саме тією галуззю NLP (обробка природної мови), яка спрямована на вирішення цієї проблеми.

Семантичний пошук – це техніка пошуку інформації, яка зосереджена на розумінні контексту та значення запитів користувача та вмісту документів для надання відповідних результатів пошуку. Алгоритм семантичного пошуку використовує методи NLP, такі як машинне навчання, глибоке навчання та онтології для вилучення й аналізу семантичних зв'язків між словами та поняттями.

Алгоритм розпочинається з попередньої обробки тексту документів і запиту користувача, що включає токенизацію, основну частину та видалення стоп-слова. Потім він ідентифікує сутності та концепції, згадані в запиті, і зіставляє їх із відповідними сутностями та концепціями в документах за допомогою різних методів, таких як розпізнавання іменованих сутностей (NER), тегування частини мови (POS) і маркування семантичної ролі (SRL).

Наступним кроком алгоритм генерує семантичне представлення запиту та документів, яке фіксує зміст і контекст тексту. Це можна зробити за допомогою таких методів, як прихований семантичний аналіз (LSA), прихований розподіл Діріхле (LDA) і нейронні мережі, такі як згорточні нейронні мережі (CNN) і рекурентні нейронні мережі (RNN).

Останній етап це ранжування документів на основі їх семантичної подібності до запиту користувача та представлення користувачеві результати

з найвищим рейтингом. Це ранжування можна зробити за допомогою різних методів, серед яких косинус подібності, подібності Жаккара та BM25.

Загалом, алгоритм семантичного пошуку – це складний процес, який включає кілька методів NLP і моделей машинного навчання, щоб надати користувачеві точні та релевантні результати пошуку.

Недоліками семантичного пошуку є те, що він досить повільний і вимагає великих обчислень.