

УДК: 004.42

Долінський А.М. аспірант, Боднарчук І.О. к.т.н., доц.

Тернопільський національний технічний університет ім. Івана Пулюя

ПОБУДОВА ЧАСТОТНОГО СЛОВНИКА ЯК ОСНОВА ЕФЕКТИВНОГО ВЕКТОРНОГО ПРЕДСТАВЛЕННЯ ТЕКСТОВИХ ДАНИХ У ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

Dolinskiy A. M. Bodnarchuk I.O.

Ternopil Ivan Puluj National Technical University

BUILDING A FREQUENCY DICTIONARY AS THE FOUNDATION FOR EFFECTIVE VECTOR REPRESENTATION OF TEXTUAL DATA IN MACHINE LEARNING TASKS

У сфері обробки природної мови (NLP) одним із ключових етапів підготовки текстових даних для машинного навчання є їх перетворення у числову форму. Найпоширенішим способом такого представлення є векторизація — тобто подання тексту у вигляді послідовності чисел, де кожне число відповідає певному слову. Для цього необхідно побудувати словник — набір унікальних слів із текстового корпусу, кожне з яких має свій числовий індекс. Ефективність навчання моделі безпосередньо залежить від якості побудови такого словника.

Найбільш ефективною технікою формування словника вважається частотний аналіз. Для цього з великого масиву текстів збирається статистика появ кожного слова. Слова, що зустрічаються надто рідко, можуть бути виключені — це дозволяє усунути шум і зменшити розмірність вхідних даних. Таким чином, до словника включаються лише найбільш поширені терміни, що мають статистичне значення для моделі.

Перед аналізом частоти слів усі тексти проходять етап попередньої обробки. Спочатку виконується токенизація — поділ тексту на окремі слова або фрази. Далі застосовується нормалізація, що включає приведення всіх символів до нижнього регістру, видалення пунктуації та спеціальних символів.

Перед аналізом частоти слова проходять попередню обробку: токенизацію, нормалізацію та спрощення через стемінг або лематизацію. Це зменшує кількість унікальних слів і покращує узагальнення та дозволяє моделі фокусуватись на змісті.

Також визначається мінімальний поріг частоти, нижче якого слова не потрапляють до словника. Це допомагає уникнути рідковживаних термінів, не зменшуючи при цьому інформаційне навантаження.

Будова частотного словника є основою ефективного векторного представлення текстів, що важливо для машинного навчання, зокрема у виявленні соціо-наукової громадської думки. Частотний аналіз дозволяє виокремити важливі терміни та зменшити шум від рідковживаних слів, що покращує навчання моделей для класифікації тенденцій у громадському дискурсі та підвищує точність прогнозів щодо соціальних настроїв.

Сформований словник використовується для векторизації, де кожне слово замінюється числовим індексом. Якщо слово відсутнє у словнику, присвоюється спеціальний індекс для невідомих термінів. У результаті кожне текстове повідомлення набуває вигляду послідовності чисел фіксованої довжини, яка використовується для навчання машинної моделі.