ПЕРЕТВОРЕННЯ СПИСКІВ ІНГРЕДІЄНТІВ СТРАВ В ОЗНАКИ (FEATURES) ДЛЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Карлов Євгеній Олександрович
студент групи ІАСм -1-24-1.4д,
Київського столичного університету імені Бориса Грінченка,
науковий керівник – к.т.н., доц. Яскевич В.О.

Списки інгредієнтів у рецептах є важливим джерелом інформації для задач класифікації рецептів, рекомендацій, персоналізації дієт, прогнозування поживності тощо. Проте вони мають кілька особливостей, які ускладнюють їхнє використання як вхідних даних для моделей машинного навчання: змінна довжина списків, неінформативний порядок компонентів, велика та розріджена словникова величина, відсутність кількісних даних або їх неоднорідність, а також наявність семантичних зв'язків між інгредієнтами (наприклад, «сметана» і «йогурт» близькі за роллю) [1].

Перш ніж кодувати інгредієнти, рекомендується виконати нормалізацію: приведення слів до нижнього регістру, усунення пунктуації, лемматизацію/стемінг, приведення інгредієнтів до уніфікованих ідентифікаторів.

Класичним підходом до кодування множини є векторне кодування Multi-hot. Кожна страва кодується бінарним вектором довжини N, де N – кількість унікальних інгредієнтів, запис "1" в векторі означає присутність інгредієнта в рецепті. Перевагами цього підходу є простота реалізації та висока інтерпретованість, недоліки – висока розмірність і розрідженість, відсутність урахування семантики/схожості між інгредієнтами.

Покращити інформативність векторного кодування можна застосувавши TF-IDF вагування, цей метод є класичним підходом у текстовій аналітиці для перетворення слів у числові ознаки. Для кожного інгредієнта і рецепту r обчислюється зважене значення $TF\text{-}IDF(i,r)=TF(i,r)\times IDF(i)$, де $TF(i,r)$ – кількість появ інгредієнта і у рецепті r (для списків інгредієнтів TF зазвичай дорівнює 1, проте можна враховувати кількість або масу інгредієнта), IDF(i) – обернена частота інгредієнта, яка зменшує вагу загальних, часто вживаних інгредієнтів. Таким чином TF-IDF зменшує вплив часто вживаних інгредієнтів та підвищує вагу рідкісних і більш інформативних. Проте кожна страва досі кодується розрідженим вектором, не враховується схожість між близькими за змістом інгредієнтами [2].

Оскільки зазвичай рецепт містить не більше кількох десятків інгредієнтів і більшість елементів закодованого вектора дорівнюють нулю, розріджені вектори можна зберігати ефективніше. Найпростіший спосіб – зберігати лише індекси ненульових елементів та їх значення, тобто кожен рецепт кодуватиметься вектором розміру кількості інгредієнтів в самому рецепті. Така оптимізація забезпечить зменшення обсягу зберігання у десятки разів.

Іншим підходом до представлення списків інгредієнтів є вектори представлення - ембеддінги (embeddings) та їх агрегація. Кожному інгредієнту

ставиться у відповідність щільний вектор (ембеддінг) визначеного розміру. Ембеддінги можна отримати в процесі навчання самої моделі або через попереднє навчання на базах рецептів. Ембеддінги зберігають семантику та дають компактні представлення. Після отримання представлень необхідно агрегувати множину інгредієнтів у фіксований вектор рецепту. Методами агрегації можуть бути просте усереднення векторів інгредієнтів (mean pooling), або вагове усереднення (IDF-weighted mean), тобто за інформативністю інгредієнта. Недоліками цього підходу є потреба в великих базах рецептів для навчання якісних ембеддінгів, втрата такими представленнями інформативності, необхідність підбору гіперпараметрів (розмір вектору представлення, метод агрегації) [3; 4; 5].

Представлення списків інгредієнтів у форматі, придатному для моделей машинного навчання, – багатогранна задача. Практичний вибір залежить від доступності даних, обчислювальних ресурсів та вимог до інтерпретованості.

ДЖЕРЕЛА

1. Diya Li, Mohammed J. Zaki, Ching-Hua Chen. Nutrition Guided Recipe Search via Pre-trained Recipe Embeddings. 2021. P. 1. URL: https://www.cs.rpi.edu/~zaki/PaperDir/DECOR21.pdf (date of access: 06.10.2025).

2. Recipe Recommendation System Using TF-IDF / Shubham Chhipa et al. 2022.
URL: https://www.researchgate.net/publication/360419221_Recipe_Recommendation_System_Using_TF-IDF (date of access: 06.10.2025).

3. Sibel Sozuer, Oded Netzer, Kriste Krstovski. A Recipe for Creating Recipes: An Ingredient Embedding Approach. 2024. URL: https://business.columbia.edu/sites/default/files-efs/citation_file_upload/WP_Recipe_Paper_01_24.pdf (date of access: 06.10.2025).

4. Abramov, V., Astafieva, M., Boiko, M., Bodnenko, D., Bushma, A., Vember, V., Hlushak, O., Zhyltsov, O., Ilich, L., Kobets, N., Kovaliuk, T., Kuchakovska, H., Lytvyn, O., Lytvyn, P., Mashkina, I., Morze, N., Nosenko, T., Proshkin, V., Radchenko, S., & Yaskevych, V. (2021). Theoretical and practical aspects of the use of mathematical methods and information technology in education and science. https://doi.org/10.28925/9720213284km

5. Карпунін, І., Зінченко, Н. Когнитивне моделювання інтелектуальних систем аналізу фінансового стану суб'єкта господарювання // Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка», 2023, 1(21), 75–85.