Инженерный вестник Дона, №7 (2022)
ivdon.ru/ru/magazine/archive/n7y2022/7780

Применение машинного обучения и анализа естественного языка для разработки рекомендательной системы по выбору парфюмерной продукции

В.Х. Смирнов, А.В. Аникин, Д.В. Литовкин, А.М. Катышев

Волгоградский государственный технический университет, Волгоград

Аннотация: В работе рассматривается использование машинного обучения применительно к обработке естественного языка (сентимент-анализа, анализа семантической близости) для построения рекомендательной системы по выбору парфюмерной продукции. Тема работы актуальна ввиду роста спектра выпускаемой парфюмерной продукции и сложности выбора её потребителями и продвижения производителями. Предлагаемые подходы релевантны для решения данной проблемы ввиду наличия накопленных текстовых отзывов и обзоров парфюмерной продукции на различных веб-сайтах, включая интернет-магазины.
Ключевые слова: машинное обучение, обработка естественного языка, сентимент-анализ, дистрибутивная семантика, word2vec, рекомендательные системы.

Введение

В настоящее время интеллектуальные рекомендательные системы используются при решении широкого спектра задач [1]. Интеллектуальные рекомендательные системы (или механизм рекомендаций) – это класс алгоритмов машинного обучения, используемых разработчиками для прогнозирования выбора пользователей и предложения соответствующих рекомендаций пользователям. Многие считают, что реализация алгоритмов рекомендаций слишком сложна и требует глобальной перестройки всего процесса сбора и обработки данных, а также изменений в бизнес-процессах и так далее. Эти сомнения необоснованны, потому что системы рекомендаций могут быть полезны практически каждому бизнесу [2], и для того, чтобы начать рекомендовать, часто достаточно уже собранных данных.

Интеллектуальные рекомендательные системы широко применяются в образовательной сфере [3], разработке различных баз знаний [4, 5], а также в решении задач из сферы маркетинга, помогая как продвигать производителям и продавцам товары и услуги до конечного пользователя,

так и пользователям с меньшими трудозатратами подбирать товары, наиболее удовлетворяющие их потребности.

Существует 4 основные типа рекомендательных систем:

- системы коллаборативной фильтрации (collaborative filtering) – рекомендации основаны на истории оценок как самого пользователя, так и других – в последнем случае в качестве входных данных рассматриваются потребители, оценки или интересы которых похожи на оценки целевого пользователя; при таком подходе, пользователи помогают друг другу в фильтрации объектов и такой метод называется также совместной фильтрацией;
- основанные на контенте (content-based) – товары и услуги рекомендуются на основе знаний об их характеристиках; в данном подходе анализируются свойства каждого объекта в системе, и на основе этих свойств тот или иной объект будет рекомендоваться определенному пользователю, у которого были найдены похожие значения анализируемых свойств объекта;
- основанные на знаниях (knowledge-based) – на основе знаний о какой-то предметной области, которые могут помочь в ранжировании: о пользователях, товарах и т.д.; у данного типа есть разделение на следующие подтипы: использование ограничений и выбор близких объектов, смысл описанных подтипов такой: у пользователей есть некоторые требования к объекту, и система пытается отыскать объект по заданным требованиям.
- гибридные (hybrid).

Специфика парфюмерной продукции подразумевает что для ее выбора не всегда достаточно знаний исключительно о базовых характеристиках продукта [6]. Таким образом, целесообразно использование гибридного подхода. С другой стороны, для сокращения расходов на сбор и анализ информации о продуктах целесообразно использование различных текстовых описаний из открытых источников (описания общих характеристик, нот

ароматов, подробных отзывов пользователей – как положительных, так и негативных) и методов анализа естественного языка для дальнейшего формирования рекомендаций [7, 8].

Подход на основе машинного обучения и анализа естественного языка для рекомендации парфюмерной продукции пользователю

В рамках работы были предложены подходы и реализованы следующие подсистемы: парсер описаний парфюмерной продукции, подсистема обучения моделей на основе анализа описаний, представленных на естественном языке, и подсистема рекомендации продукции пользователю на основе его предпочтений.

Парсер описаний парфюмерной продукции

Парсер описаний парфюмерной продукции из открытых источников на примере одного из интернет-магазинов. Реализован на языке Python с использованием библиотеки BeautifulSoup. Beautiful Soup – это библиотека Python для извлечения данных из HTML- и XML-файлов. Данная библиотека работает с любым синтаксическим анализатором, предоставляя удобные способы навигации, поиска и изменения дерева синтаксического анализа.

В результате дальнейшего использования парсера был собран датасет [9] из 3900 наименований парфюмерной продукции, включающий следующие поля: наименование, изображение, вид продукции, ноты аромата (верхние, сердечные, шлейфовые), массив отзывов покупателей.

Подсистема обучения модели

С использованием сентимент-анализатора VADER отзывы покупателей в датасете были проанализированы, положительные и нейтральные отзывы были объединены в одну строку, все отрицательные отзывы были объединены в другую строку.

Остальные поля записей в датасете были объединены в отдельную строку для каждой записи и использовались в качестве документов для обучения эмбединг-моделей с использованием двух подходов: LSA (латентно-семантический анализ) и Doc2Vec [10] (подход векторизации текстовых документов на основе эмбедингов отдельных слов документа).

LSA использует токенизацию слов в документе на основе подхода TF-IDF, после чего сжимает набор полученных характеристик с использованием сингулярного разложения (SVD), реализуя таким образом модель «мешок слов» (Bag-of-Words, BOW), в которой контекст использования слов и их последовательность в тексте не учитываются. Тем не менее, данный подход является одним из классических и успешно показывает себя во многих задачах анализа естественного языка.

Doc2Vec является нейросетевым подходом анализа документов на естественном языке на основе обучения эмбедингов из текстового документа. Модель Doc2Vec используется для создания векторного представления группы слов, взятых вместе как единое целое. Алгоритм данной модели использует модель нейронной сети для изучения ассоциаций слов из большого массива текста. После обучения такая модель может обнаруживать синонимичные слова или предлагать дополнительные слова для частичного предложения. Doc2Vec представляет каждое отдельное слово с определенным списком чисел, называемым вектором. Векторы выбираются тщательно таким образом, чтобы простая математическая функция (косинусное сходство между векторами) указывала уровень семантического сходства между словами, представленными этими векторами. Таким образом, в отличие от предыдущего подхода, данная модель учитывает контекст слов, возможные связи между словами, их последовательности для моделирования возможного смысла текста (предложения или документа), что важно при построении рекомендательной системы парфюмерной продукции, т.к. оценка

продукта и пожелания к нему в большинстве случаев не выражаются точными значениями набора его характеристик.

Таким образом, целесообразно комбинирование обеих подходов для повышения качества модели и результатов работы рекомендательной системы.

Подсистема генерации рекомендаций на основе предпочтений пользователя

В качестве входных данных подсистемы используется обученная модель, а также строка текста, задаваемая пользователем, которая может включать как непосредственно отдельные желаемые характеристики продукции, так и более сложные описания, не относящиеся напрямую к данным характеристикам (настроение, предполагаемое использование продукта – место или событие (торжество, вечер, офис), стойкость, ассоциации с ароматом и другие). Подсистема вычисляет векторное представление введенного запроса на основе подходов LSA и Doc2Vec, затем производится сентиментный анализ предложений запроса и вычисляется косинусная близость векторов положительных и нейтральных предложений запроса к документам датасета, описывающим продукты, с дальнейшей фильтрацией тех документов, которые имеют высокую косинусную близость к векторам негативных предложений запроса. Косинусная близость вычисляется отдельно с использованием подходов LSA и Doc2Vec, после чего показатели усредняются.

Заключение

В работе рассмотрен подход на основе машинного обучения и анализа естественного языка (дистрибутивной семантики, сентимент-анализа) для разработки интеллектуальной рекомендательной системы по выбору

парфюмерной продукции на основе анализа корпуса текстов описания продукции и отзывов покупателей.

В дальнейшем развитии работы планируется расширение корпуса описаний и отзывов о парфюмерной продукции на основе различных открытых источников для повышения качества модели, тестирование модели с определением соответствующих показателей качества (точность, полнота, F-мера, ROC-AUC (площадь под кривой ошибок)), реализация веб-сервиса рекомендации парфюмерной продукции и внедрение в коммерческую эксплуатацию.

Литература (References)

1. Fanca A., Puscasiu A., Gota D., Valean H. Recommendation Systems with Machine Learning. 2020 21th International Carpathian Control Conference (ICCC). IEEE, 2020. DOI: 10.1109/iccc49264.2020.9257290.

2. Anikin A., Katyshev A., Denisov M., Smirnov V., Litovkin D. Using Online Update of Distributional Semantics Models for Desicion-Making Support for Concepts Extraction in the Domain Ontology Learning Task. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012073.

3. Anikin A., Litovkin D., Sarkisova E., Petrova T., Kultsova M. Ontology-based approach to decision-making support of conceptual domain models creating and using in learning and scientific research. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012074.

4. dos Santos, P.V., Kuehne, B.T., Batista, B.G., Leite, D.M., Peixoto, M.L., Moreira, E.M. and Reiff-Marganiec, S. Recommender Systems Evaluator: A Framework for Evaluating the Performance of Recommender Systems. In: Shahram, L. (Ed.) ITNG 2021 18th International Conference on Information Technology-New Generations, New York: Springer, 2021, pp. 339-345. DOI: 10.1007/978-3-030-70416-2_43.

5. Singh Pramod. Recommender Systems. Machine Learning with PySpark. Apress, Berkeley, CA, 2022. pp.157-158. DOI: 10.1007/978-1-4842-7777-5_8.

6. Learning to Smell: Using Deep Learning to Predict the Olfactory Properties of Molecules. Google AI Blog URL: ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html (access date: 01.12.2021).

7. Hanafizadeh Payam, Ravasan Ahad Zare, Khaki Hesam Ramazanpour. An expert system for perfume selection using artificial neural network. Expert Systems with Applications, 2010, № 37, p. 8879-8887.

8. Danli Wu, Cheng Yu, Luo Dehan, Wong Kin-Yeung, Hung Kevin, Yang Zhijing. POP-CNN: Predicting Odor's Pleasantness with Convolutional Neural Network. URL: arxiv.org/ftp/arxiv/papers/1903/1903.07821.pdf (access date: 01.12.2021).

9. Bansal Nandini. Perfume Recommendation Dataset. URL: kaggle.com/nandini1999/perfume-recommendation-dataset (access date: 15.12.2021).

10. Le Quoc, Mikolov Tomas. Distributed Representations of Sentences and Documents. URL: cs.stanford.edu/~quocle/paragraph_vector.pdf (access date: 15.12.2021).