«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Центр непрерывного образования

Факультета компьютерных наук

ИТОГОВЫЙ ПРОЕКТ

СОПОСТАВЛЕНИЕ НАЗВАНИЙ ТОВАРОВ ИЗ АССОРТИМЕНТА АПТЕК

Выполнил (а):

Конюхова Мария Вячеславовна

Ф.И.О.

Руководитель:

Арк Михаил Юрьевич

Ф.И.О.

Москва 2024

Оглавление

I. Введение

В современном мире, где объемы данных растут с каждым днем, возникает необходимость в их структурировании и анализе. Сопоставление товаров – является одной из важнейших задач при принятии решений, особенно в таких отраслях, как электронная коммерция, управление цепочками поставок и розничная торговля. Использование алгоритмов сопоставления помогает повышать эффективность управления запасами, а также помогает клиентам находить нужные товары у поставщиков, сравнивать цены и принимать обоснованные решения о покупке. Также сопоставление товаров позволяет проводит анализ продуктовой стратегии («что есть у меня и нет у конкурентов» и наоборот).

Однако, если говорить про фармацевтическую отрасль, разнообразие наименований, форм выпуска, дозировок создает сложности для точного сопоставления товаров, что в свою очередь может привести к ошибкам в заказах, учете и, что наиболее важно, к серьезным последствиям и риску для здоровья пациентов (в случае ошибочной поставки жизненно важного лекарства). Использование алгоритмов сопоставления не только повышает эффективность управления запасами, но и позволяет обеспечить более высокий уровень безопасности и обслуживания клиентов. Важность задачи обусловлена также и экономической составляющей, поскольку сопоставление товаров может значительно снизить издержки и увеличить прибыльность бизнеса (при сопоставлении цен товара у разных поставщиков и изменении ценовой политики).

Целью работы является разработка алгоритма, способного с достаточно высокой точностью классифицировать товары и осуществлять сопоставление наименований товаров из ассортимента аптеки со списком товаров поставщиков. В качестве исходной информации был получен массив данных, содержащий совокупность данных об аптечных товарах от разных

поставщиков. В рамках работы для каждой строчки из списка аптеки осуществлялся подбор подходящих товаров из списка поставщика.

В связи с этим, можно выделить две основных задачи работы:

1) Провести классификацию для товаров из списка аптеки и для товаров из списка поставщика, используя разные алгоритмы машинного обучения, при этом оценив их результативность.
2) Оценить эффективность сопоставления товаров путем расчета бизнес-метрики «процент сопоставлений» (процент совпадений категорий (классов) товаров из списка аптеки и списка поставщика по предсказанным категориям).

Кроме того, в работе дополнительно были рассмотрены задачи реализации алгоритма сопоставления товаров с использованием векторизации на основе TF-IDF (Term Frequency-Inverse Document Frequency) и расчета косинусного сходства для определения степени соответствия между товарами, а также вариант реализации сопоставления с использованием нормализованного расстояния Левенштейна.

Для достижения вышеуказанных целей и задач была изучена литература по теме, проведена предобработка и подготовка данных, поставлены эксперименты в части классификации с разными алгоритмами машинного обучения, использования косинусного сходства и расстояния Левенштейна при осуществлении сопоставления, проведена оценка результативности. Работа проводилась в июне 2024 года.

II. Обзор литературы

С большим ростом объема данных появляются новые задачи машинного обучения и новые способы их решения. Ранее задача сопоставления уже проявлялась как в поисках дубликатов, так и в классических приложениях по восстановлению целостности данных. Последние несколько лет появились возможности проводить более сложные семантические сопоставления неструктурированных данных, таких как фото, видео, текст. Использование тех или иных методов сопоставления товаров в первую очередь зависит от наличия и качества исходных данных для сопоставления и целей сопоставления, соответственно.

В настоящее время на рынке существует множество компаний и сервисов, предлагающие услуги автоматического и ручного сопоставления товаров по различным требованиям, например, автоматизированная платформа мониторинга цен и ценообразования Priceva (https://priceva.ru/), сервисы мониторинга данных MarketParser (https://marketparser.ru/), AllRival (https://allrival.com/), iDatica (https://idatica.com/) и другие.

Сфера обработки естественного языка переживает бурный рост благодаря развитию технологий машинного обучения. Среди прочих категорий задач в Natural Language Processing (NLP) классификация текстов является одной из наиболее часто встречающихся. В последние годы было проведено множество исследований, посвященных использованию классификационных алгоритмов для сопоставления товаров, и постоянно осуществляется оптимизация и совершенствование процессов. Часто в задачах классификации текста применяется TfidfVectorizer - инструмент для преобразования текстовых данных в числовые признаки, которые затем используются для обучения моделей машинного обучения. Алгоритмы машинного обучения, такие как деревья решений, случайный лес, и другие могут быть использованы с целью классификации после того, как текст был преобразован в числовые признаки.

Рассмотрение различных источников информации позволяет выявить ключевые подходы и методы, используемые для решения поставленных задач.

При осуществлении работы над поставленными задачами проведено ознакомление с базовой теорией и документацией по использованию алгоритмов машинного обучения, описанных в библиотеке Scikit-Learn [1], в материалах семинаров Е. Волошиной [2], в статье А. Ланского [3], Y. Kashnitsky [4], в образовательном блоге о языке программирования Python [5], в материалах курса по машинному обучению Е. Соколова [6], в статье Е. Лабинцева [7] и материалах сайта А. Мичурина [8] о метриках в задачах машинного обучения, многочисленных видеоматериалах интернет-портала youtube.com и статьях интернет-портала habr.ru.

Кроме того, были рассмотрены подходы к классификации текста, описанные в работах А. Старченкова [9], K. Dynev [10], в видеолекции о многоклассовой классификации текста А. Галямова [11], в магистерской диссертации Abiola A.David [12], в статье В. Дарморезова [13].

Необходимо отметить, что довольно часто встречается решение задачи сопоставления с помощью расчета и анализа косинусного сходства и расстояния. В этой части хотелось бы отметить статью А. Ткаченко [14], проект Retail Product NLP-Match Cluster [15], проект Product matching [16], подходы, предложенные в рамках видеоинтервью В. Бабушкина [17]. Также решение задачи часто осуществляется с помощью расчета и анализа расстояния Левенштейна, например, с использованием библиотеки TheFuzz [18] и подхода, описанного в статье Ezequiel Ortiz Recalde [19].

III. Методы

Необходимо отметить, что наименование товара у разных поставщиков сформулировано по-разному (без четких требований и норм), но указанная в массиве данных категория товара (класс) является стандартным наименованием товара, которое мы использовали как целевую переменную при осуществлении классификации.

Для решения задачи классификации в работе были рассмотрены следующие методы (с рассмотрением для каждого из них метрик accuracy, precision, recall, F1-score):

- наивный байесовский классификатор;
- дерево решений;
- случайный лес;
- KNN;
- метод опорных векторов.

Оценка качества сопоставления товаров проводилась на основании бизнес-метрики «процент сопоставлений», рассчитанной как доля товаров, предсказанные категории по которым по списку аптеки и списку поставщика совпали. Также в дополнение после проведения классификации сформирован список всех имеющихся поставщиков из исходного датасета на основании предсказанных категорий (DecisionTreeClassifier) из списка товаров аптеки с использованием метода объединения данных (data merging) и группировки (grouping) (с целью извлечения уникальных значений поставщиков по категории товара), что с практической точки зрения может быть применимо при определении поставщиков, с которыми предполагается сотрудничество.

Кроме того, в работе был рассмотрен вариант сопоставления списков товаров с использованием метода TF-IDF (Term Frequency-Inverse Document Frequency) в сочетании с косинусным сходством. При решении задачи данным методом сопоставление осуществлялось по сформированным спискам товаров, наименования которых представлялись в виде векторов, где каждое

слово является измерением. Вычислив косинусное сходство между двумя векторами произведений, количественно оценивалось их текстовое сходство.

Также в работе был рассмотрен вариант реализации задачи сопоставления с использованием нормализованного расстояния Левенштейна, которое позволяет учесть различия в длине строк.

Результаты сопоставления (и в случае с косинусным сходством, и в случае с расстоянием Левенштейна) оценивались в зависимости от различных пороговых значений с учетом метрик precision, recall, F1-score.

IV. Эксперименты

Перед тем, как начать работать с алгоритмами классификации, была проведена предообработка данных в части удаления нелекарственных товаров из массива данных, дозаполнения колонки наименования поставщика с использованием вспомогательной колонки, где также был указан поставщик, а также удаления строк, содержащих пустые значения в колонках наименования товара, поставщика, категории товара. Кроме того, была проведена предобработка строк в части преобразования текста в нижний регистр, удаления лишних пробелов, стоп-слов (в т.ч. слов «скидка», «подарок», «акция», «внимание», «неопознанный»), замены всех знаков препинания на знак «точки» и сокращение всех последовательностей «точек» до одной.

Далее был организован процесс выборки из массива данных 1 500 уникальных категорий (рандомно), по каждой из которых были взяты все имеющиеся к ним наименования товаров (с условием – от 4 до 20 наименований на одну категорию). Впоследствии было проведено разделение выбранного массива на три части: первая строка по каждой категории данной выборки была отнесена в список аптеки, вторая – в список поставщика, все остальные строки для той же категории товара – в обучающую выборку. Кроме того, был создан словарь «идеала сопоставления», «ключом» в котором являлось наименование товара из списка аптеки, а «значением» – наименование товара из списка поставщика.

Соответственно, проведено разбиение данных на обучающую и тестовые выборки. В нашем случае обучающая выборка составила 24 000 строк, в качестве тестовых выборок использовались список товаров аптеки (1 500 строк) и список товаров поставщика (1 500 строк), как уже было отмечено выше.

Классификация производилась пятью методами (наивный байесовский классификатор, дерево решений, случайный лес, KNN, метод опорных векторов).

Классификаторы оценивались на тестовой части выборки (не используемой при обучении) с помощью базовых метрик качества – accuracy, precision, recall, F1-score. Также в каждом варианте рассчитывалась бизнес-метрика «процент сопоставлений» (доля товаров, предсказанные категории по которым по списку аптеки и списку поставщика совпали). Также в работе использовались модули time и psutil с целью мониторинга и анализа производительности, эта информация также учитывалась при формировании выводов.

После обучения алгоритмы показали результаты, представленные в таблице:

1 500 категорий (обучающая выборка 24 000 строк)		NBC	DT	RF	SVM	KNN
		3 сек.	490 сек.	370 сек.	196 сек.	1 сек.
		207 мб	59 мб	7 112 мб	464 мб	11 мб
список аптеки	accuracy	0,87	0,96	0,96	0,95	0,93
	precision	0,81	0,95	0,95	0,94	0,9
	recall	0,87	0,96	0,96	0,95	0,93
	F1-score	0,83	0,95	0,95	0,94	0,91
список поставщика	accuracy	0,87	0,96	0,96	0,96	0,94
	precision	0,81	0,95	0,95	0,94	0,92
	recall	0,87	0,96	0,96	0,96	0,94
	F1-score	0,83	0,95	0,95	0,95	0,92
процент сопоставлений		0,86	0,95	0,95	0,94	0,92

На основании этого можно сделать вывод, что высокие метрики получаются с алгоритмами DT и RF. Однако случайный лес потенциально

может быть медленным и дорогостоящим в вычислительном отношении, особенно для больших наборов данных или при использовании большого количества деревьев, поэтому его использование для целей нашей работы не рекомендуется. Наиболее оптимальным вариантом из предложенных является использование DT (если отсутствуют жесткие требования по скорости выполнения задачи) или SVM. Однако необходимо отметить, что подобная многоклассовая классификация довольно трудоемка, и задача может быть решена оптимальнее иными способами, существующими в отрасли.

В рамках реализации метода ближайших соседей KNN был произведен подбор параметра n_neighbors с помощью кросс-валидации – получено оптимальное значение k=5. Во всех алгоритмах обучение было проведено с параметрами «по умолчанию». Но для более качественной работы моделей конечно же лучше проводить их дополнительную настройку.

По алгоритму с наиболее высокими метриками (DT) был произведен подбор поставщиков из базы данных поставщиков по списку товаров аптеки, результаты которого можно использовать для определения поставщика (или поставщиков) с которыми предполагается сотрудничество в части заказа необходимых аптеке товаров.

В части реализации сопоставления списков товаров с использованием метода TF-IDF в сочетании с косинусным сходством использовались список аптеки и список поставщика, ранее сформированные на начальном этапе классификации. По предобработанным данным была проведена TF-IDF векторизация, включающая объединение текстов из обоих массивов данных и их преобразование в векторное представление, получение матриц для каждого списка. Затем произведено вычисление косинусного сходства между всеми парами из двух списков (пары считались сопоставленными, если их косинусовое сходство превышало пороговое значение). В нашем случае было рассмотрено несколько вариантов размера порога косинусного сходства и проведена оценка базовых метрик precision, recall, F1-score:

Изменение порога	0,3	0,4	0,5	0,6	0,7
precision	0.3692	0.6628	0.8464	0.9037	0.9448
recall	0.9020	0.7927	0.6393	0.4567	0.2967
F1-score	0.5239	0.7219	0.7284	0.6067	0.4515

По мере увеличения порога наблюдается постепенное снижение recall и увеличение precision. Это ожидаемое поведение, поскольку при увеличении порога мы становимся более строгими в отборе пар наименований. При низких порогах (0,1 и 0,2) отмечена высокая полнота (recall) и низкая точность (precision), что указывает на большое количество ложных срабатываний. При высоких порогах (0,8 и 0,9) – высокая точность и низкая полнота, что указывает на то, что модель пропускает многие истинные пары, но среди найденных пар почти все являются правильными. Средние пороги (0,4 и 0,5) дают лучший баланс между точностью и полнотой, что отражается в высоком значении F1-метрики.

При пороге в 0,4 точность составляет примерно 66,28%, что указывает на то, что примерно две трети классифицированных совпадений действительно корректны. Полнота достигает примерно 79,27%, что означает, что система смогла обнаружить около 80% всех потенциально правильных совпадений. F1-score, которая является гармоническим средним между точностью и полнотой, составляет 72,19%, что свидетельствует о хорошем балансе между этими двумя метриками.

При более высоком пороге в 0,5 точность возрастает до 84,64%, показывая, что большинство классифицированных совпадений являются верными. Однако, полнота снижается до 63,93%, что означает, что система находит меньше правильных совпадений из всех возможных. F1-score при этом пороге составляет 72,84%, что также указывает на достаточно сбалансированное соотношение между точностью и полнотой, несмотря на изменения в этих двух метриках.

Кроме того, был рассмотрен подход к реализации задачи сопоставления с использованием нормализованного расстояния Левенштейна, которое позволяет учесть различия в длине строк. Для этого была применена функция из библиотеки TheFuzzy. Было произведено вычисление нормализованного расстояния Левенштейна между каждой парой названий товаров, чтобы определить степень их сходства. Затем, на основе выбранного порогового значения, пары сопоставлялись как схожие или различные. Сопоставление также проводилось с использованием различных пороговых значений. Для каждого порога были рассчитаны базовые метрики, такие как precision, recall и F1-score. Перебор пороговых значений позволил определить оптимальный порог для максимизации точности сопоставления:

Изменение порога	0,2	0,3	0,4	0,5	0,6
precision	0.9111	0.6572	0.1943	0.0333	0.0044
recall	0.1913	0.4013	0.6693	0.8793	0.9680
F1-score	0.3163	0.4983	0.3012	0.0641	0.0088

Наиболее оптимальным выглядит порог в 0,3. При пороге 0,3 точность (precision) составляет примерно 65,7%, что означает, что из всех найденных совпадений около двух третей действительно являются правильными. Полнота (recall) составляет примерно 40,1%, что означает, что было найдено только около 40% всех возможных правильных совпадений. F1-score составляет примерно 49,8%, что является довольно сбалансированным показателем, учитывая компромисс между точностью и полнотой. Однако необходимо отметить, что в сравнении с ранее рассмотренным методом согласно полученным метрикам для наших данных предпочтительнее вариант TF-IDF с оценкой косинусного сходства (по скорости реализации он также оптимальнее).

V. Заключение

В процессе работы были проведены эксперименты и получены следующие результаты:

1. Можно сделать вывод, что для целей сопоставления товаров из разных списков можно использовать рассмотренные варианты (классификация с разными алгоритмами машинного обучения, использование TF-IDF совместно с косинусным сходством и подход с оценкой расстояния Левенштейна). Каждый из рассмотренных вариантов обладает своими особенностями и метриками качества.
2. В части классификации крайне не рекомендуется алгоритм RF ввиду низкой скорости реализации и высокой стоимости в вычислительном отношении. Более предпочтительными являются алгоритмы DT и SVM. Однако подобная многоклассовая классификация довольно трудоемка, и задача может быть решена оптимальнее иными способами, существующими в отрасли.
3. Сопоставление списков товаров с использованием метода TF-IDF в сочетании с косинусным сходством для наших данных является предпочтительнее метода с использованием нормализованного расстояния Левенштейна. При пороге в 0,4 примерно две трети классифицированных совпадений действительно корректны, система смогла обнаружить около 80% всех потенциально правильных совпадений. При более высоком пороге в 0,5 точность возрастает до 84,64%, показывая, что большинство классифицированных совпадений являются верными, однако система находит меньше правильных совпадений из всех возможных.

4. С учетом имеющихся данных написан код для их обработки. Информацию можно найти в открытом доступе в репозитории проекта[1].

Улучшение моделей — постоянная задача. Если производительность модели неудовлетворительна, можно использовать различные стратегии для повышения ее точности. Эти стратегии включают в себя сбор более размеченных данных, проведение более тщательной предобработки, настройку гиперпараметров, разработку новых функций или использование более продвинутых и оптимальных методов решения задачи сопоставления текста.

[1] https://github.com/maffka-brox/Product-Matching/tree/main

Список литературы

1) Библиотека Scikit-Learn. https://scikit-learn.org/stable/index.html.
2) Е. Волошина. Материалы семинаров; 2022. https://github.com/hse-ling-python/seminars/tree/master/ml
3) А. Ланский. Обзор методов классификации в машинном обучении с помощью Scikit-Learn; 2019. https://tproger.ru/translations/scikit-learn-in-python
4) Y. Kashnitsky. ML Course.AI; 2023. https://github.com/Yorko/mlcourse.ai/tree/main/mlcourse_ai_jupyter_book/book
5) Образовательный блог о языке программирования Python. https://pythonru.com/
6) Е. Соколов. Материалы курса по машинному обучению; 2016. https://github.com/esokolov/ml-course-msu/tree/master/ML15/lecture-notes
7) Е. Лабинцев. Метрики в задачах машинного обучения; 2017. https://habr.com/ru/companies/ods/articles/328372/
8) А. Мичурин. Интернет-портал. https://www.michurin.net/computer-science/ml-precision-recall.html
9) А. Старченков. Решаем NLP-задачу классификация текстов по темам; 2022. https://dzen.ru/a/Yk7qXIk6T1a25vxD
10) K. Dynev. Создание модели предсказания кода МКБ-10 на основе текста описания болезни; 2022. https://habr.com/ru/articles/673312/
11) А. Галямов. Видеолекция «Многоклассовая классификация текста на Python»; 2020. https://www.youtube.com/watch?v=iNl6-3KvpHk&ab_channel=%D0%90%D0%B9%D1%80%D0%B0%D1%82%D0%93%D0%B0%D0%BB%D1%8F%D0%BC%D0%BE%D0%B2
12) Abiola A. David. Master's thesis «Product Matching: A Comparative Analysis of Various Machine Learning Algorithms using Word2Vex and TF-IDF Embedding Techniques»; 2024.

https://www.linkedin.com/pulse/product-matching-comparative-analysis-various-machine-abiola-hpene/

13) В. Дарморезов. Пайплайн для создания классификации текстовой информации; 2023. https://habr.com/ru/articles/724790/

14) А. Ткаченко. Сравнение текста с помощью статистической меры TF-IDF; 2020г. https://newtechaudit.ru/kak-sverit-tekst-tf-idf/

15) P. R. Lopez. Retail Product NLP-Match Cluster; 2024. https://github.com/pablo-git8/RetailProductNLP-MatchCluster

16) Data Rizzz. Product matching; 2022. https://www.kaggle.com/code/sohamdas27/product-matching/notebook

17) В. Бабушкин. ML System Design с Валерием Бабушкиным, выпуск 3, собеседование, karpov.courses; 2022. https://www.youtube.com/watch?si=YdhsR4cJ3OfWhyda&v=3X-TAuWdIAc&feature=youtu.be

18) Библиотека TheFuzz. https://github.com/seatgeek/thefuzz

19) Ezequiel Ortiz Recalde. A Fuzzy String Matching Story; 2021. https://towardsdatascience.com/a-fuzzy-string-matching-story-314bbecaa098