Классификация товаров по ОКПД2 кодам

A Preprint

Фирсов Сергей
Кафедра интеллектуальных систем
МФТИ
firsov.sa@phystech.edu

Всеволод Михайлович Старожилец
Кафедра интеллектуальных систем
Форексис
vsevolod.starozhilets@antirutina.net

Abstract

Исследование направлено на решение задачи классификации товаров по ОКПД 2 кодам с использованием кратких текстовых описаний. Коды представляют собой детализированную систему категоризации продуктов и услуг по видам экономической деятельности. Основная цель - повышение точности и сокращение ресурсозатратности классификации, анализируя влияние глубины классификатора ОКПД 2. Для достижения этих целей предлагается метод построения текстовых эмбеддингов с использованием нейросетевых технологий, таких как spaCy. Задача усложняется необходимостью предварительной обработки данных для перевода исходных описаний в стандартизированные короткие тексты, адаптированные для анализа. Используются данные государственных закупок по ФЗ 44 за 2022 год, охватывающие около 40% открытых источников, что обеспечивает достаточный объем и разнообразие информации для анализа. Также часть этих данных будет критерием оценки работы программы. Новизна заключается в применении методов машинного обучения к индустриальной задаче, что обещает улучшение в процессах логистики, учёте и анализе в сфере закупок.

Keywords OKPD 2 code · text analysis · task of classification

1 Введение

Целью данного исследования является разработка и апробация метода классификации товаров по ОКПД 2, используя краткие текстовые описания. Актуальность задачи обусловлена необходимостью повышения эффективности процессов логистики и учета в сфере закупок, а также сокращения времени и ресурсов, затрачиваемых на классификацию товаров.

Объектом исследования выступают любые товары, для которых возможна классификация по ОКПД 2 кодам (детализированной системе категоризации продукции и услуг по видам экономической деятельности). Проблема заключается в разработке метода, позволяющего автоматизировать этот процесс с высокой точностью и полнотой классификации, устойчиво относительно формата входных данных, и в исследовании характеристик этого метода (по указанным параметрам) от глубины классификации.

Задача классификации разобрана вдоль и поперёк в любой доступной литературе по машинному обучению. В дальнейшей работе будем опираться на курс лекций Воронцова К.В. и книги Гудфеллоу et al. [2016] и Montani [2019]. Текстовые эмбеддинги тоже активно используются в современной разработке, есть много предобученных моделей и пакетов, таких как Word2Vec, GloVe, spaCy, по ним также найдены и частично изучены книги и современные результаты.

Предлагаемое решение базируется на использовании передовых методов обработки естественного языка и машинного обучения: получения словаря слов-признаков на основе данных госзакупок, построение эмбендингов по ним, с использованием spaCy и решение задачи классификации по полученным векторам на основе результатов обучающей выборки. Решение ново настолько, насколько нов подход к решению индустриальной задачи с помощью методов машинного обучения. Преимуществами такого подхода

являются увеличение точности и снижение затрат времени на обработку данных, как раз то что требуется в прикладных задачах.

Цель эксперимента – проверка эффективности предложенного метода на реальных данных государственных закупок за 2022 год. Это позволяет оценить работу алгоритма в условиях большого объема и разнообразия данных. Экспериментальная установка включает в себя подготовку данных, построение модели классификации и ее тестирование с целью определения оптимальных параметров для достижения максимальной точности классификации в зависимости от необходимой глубины класификации. Рабочий процесс описывает последовательные шаги от предварительной обработки текстов до оценки результатов классификации.

В заключение, данное исследование представляет собой вклад в развитие методов машинного обучения и их применение к решению практических задач классификации товаров, что имеет важное значение для сферы государственных закупок и управления цепочками поставок.

2 Постановка проблемы

Sample set (X) - данные госзакупок за 2022 год. Представляют собой таблицу из двух колонок, в первой - краткое (почти наверное) описание товара, во второй - заданный ОКПД 2 код этого товара. Выборка из более 8 миллионов результатов, считается полностью корректной, с точностью до описания товаров. Записи могут частично повторяться или незначительно отличаться текстовым описанием.

Будет использоваться линейная модель классификации, $A = \{ g(x,\theta) \mid \theta \in \mathcal{R} \}$, где $g(x,\theta) = sign \sum_{j=1}^{n} \theta_j f_j(x)$.

Используется квадратичная функция потерь: $\mathcal{L}(a,x) = (a(x) - y(x))^2$, где y - значения контрольной выборки.

Откуда получаем функцию Эмпирического риска : $Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(a, x_i)$

И будем решать задачу оптимизации - минимизации эмпирического риска

$$\mu(X^l) = arg \min_{a \in A} Q(a, X^l)$$

Критерий качество в задаче нашей классификации предельно ясен - попадание конкретного товар в свою категорию и в правильный ОКПД 2 код.

Список литературы

Ян Гудфеллоу, Йошуа Бенджио, and Аарон Курвилль. Глубокое обучение. MIT Press, 2016.

Ines Montani. Advanced NLP with spaCy: A Practical Guide to Advanced Natural Language Processing. Independent, 2019.

1. Лучше расписать что такое ОКПД один раз и в скобках поставить сокращение
2. по всему тексту - здесь нужно писать не дефис, а длинное тире. В латехе пишется как три дефиса "---"
3. "Повышение точности, анализируя" - части предложения несогласованы. Нужно переписать
4. Здесь не имеет смысла указывать библиотеку, которая используется. Ничего в вашем методе не поменяется, если вы возьмете вместо spacy другую библиотеку.
5. Очень много деталей, которые читателю аннотации не будут интересны. 40% убрать, ссылку на ФЗ - возможно тоже (уточните у Всеволода).
6. Выборка не является критерием (может, качество модели на выборке?). Желательно писать аннотацию в настоящем времени
7. Новизна чего? Работы? Метода? Уточнить
8. Можно добавить ссылку на классификатор
9. Так не пишут. Укажите основые ссылки на работы, схожие с вашей, и описание этих работ. По классификации кодов работ достаточно. Если не найдете - дайте знать, подкину. (лучше на опубликованные работы, курс лекций Воронцова замечательный, но не опубликован, к сожалению)
10. Если вы используете эмбеддинги, то такого описания также недостаточно. Напишите по одному-два предложения про каждую работу. Также нужно объяснить что такое эмбеддинг. Как вариант - писать "векторное предложение слова"
11. Передовые - понятие относительное. Опишите словами что это за методы. Само слово "передовые" нужно убрать
12. см. комментарий 4
13. Это предложение неинформативно и не нужно и не соответствует научному стилю статьи
14. Эта часть предложения неинформативна, не нужна и не соответствует научному стилю статьи
15. Тут нужны вводные слова. Читателю до этого предложения надо сообщить, что в работе проводится эксперимент
16. Посмотрите на работы с прошлого года. Так не пишут. Введите выборку, опишите, из каких частей она состоит (видимо, текстовое описание и метки). Какому множеству принадлежат метки, как представляется выборка с помощью эмбеддингов
17. Выборка не состоит из двух колонок, выборка может состоять из двух математических объектов (условно, X и y)
18. Обсудите с Всеволодом математическиую нотацию. Мы на курсе предлагаем следующее:
    * векторы и вектор-функции - жирные прописные буквы
    * матрицы - заглавные жирные буквы

(см. слайды с предыдущего занятия)
19. Формально формулы являются частями предложений, поэтому после каждой формулы должны стоять запятые или точки.
20. Критерий качествА. И лучше написать его формально, по формуле.