Евстифеев И.А. Перспективы использования машинного обучения для классификации документов государственных закупок // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 8 № 6(32) с. 28–39

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ ГОСУДАРСТВЕННЫХ ЗАКУПОК

Евстифеев И.А.
ФГБОУ ВО Самарский государственный технический университет, Самара, Россия (443010, г. Самара, Молодогвардейская ул., 244.), e-mail: il.evs@yandex.ru

В данной статье будет произведена попытка разработки алгоритма, который автоматически будет классифицировать документы государственных закупок по ОКПД2 кодам. Это позволит значительно упростить и ускорить процесс анализа закупок, а также повысить эффективность мониторинга государственных закупок. Для достижения этой цели здесь будут использоваться методы машинного обучения, алгоритмы классификации. Так же будет проведен анализ различных методов предобработки данных, чтобы определить наиболее эффективные способы обработки информации. В итоге, результаты этой работы могут быть применены в различных областях, связанных с государственными закупками, включая анализ и оптимизацию бюджетных расходов, мониторинг конкуренции на рынке, а также принятие решений при выборе поставщиков товаров и услуг.

Ключевые слова: алгоритмы классификации, государственные закупки, ОКПД2 код, методы предобработки данных.

PROSPECTS FOR USING MACHINE LEARNING TO CLASSIFY PUBLIC PROCUREMENT DOCUMENTS

Evstifeev I.A.
Samara State Technical University, Samara, Russia (443010, Samara, Molodogvardeyskaya str., 244), e-mail: il.evs@yandex.ru

In this article, an attempt will be made to develop an algorithm that will automatically classify public procurement documents by OKPD2 codes. This will significantly simplify and speed up the procurement analysis process, as well as improve the effectiveness of public procurement monitoring. To achieve this goal, machine learning methods and classification algorithms will be used here. The analysis of various data preprocessing methods will also be carried out to determine the most effective ways of processing information. As a result, the results of this work can be applied in various areas related to public procurement, including analysis and optimization of budget expenditures, monitoring of competition in the market, as well as decision-making when choosing suppliers of goods and services.

Keywords: classification algorithms, public procurement, OKPD2 code, data preprocessing methods.

Цель работы

Цель данной работы заключается в разработке алгоритма классификации документов государственных закупок по видам экономической деятельности с использованием Общероссийского классификатора продукции по видам экономической деятельности (ОКПД2) кодом. Это позволит автоматизировать и упростить процесс анализа закупок, а также повысить эффективность мониторинга государственных закупок. Для достижения этой цели в

работе будут использоваться методы машинного обучения, включая нейронные сети и алгоритмы классификации, а также проводиться анализ различных методов предобработки данных, чтобы определить наиболее эффективные способы обработки информации. Результаты работы могут быть применены в различных областях, связанных с государственными закупками.

Задачи

1. Сбор и анализ данных о государственных закупках, включая информацию о товарах и услугах, их стоимости, дате заключения контрактов, наименовании заказчиков и исполнителей.
2. Предобработка данных, включая очистку, нормализацию, токенизацию и лемматизацию текстовых данных, а также преобразование числовых данных в удобный для анализа формат.
3. Построение модели классификации, включая выбор подходящих алгоритмов и методов машинного обучения, настройку гиперпараметров и оценку качества модели.
4. Анализ результатов классификации, включая оценку точности, полноты, F-меры и других метрик качества, а также исследование влияния различных методов предобработки данных на результаты классификации.
5. Разработка программного обеспечения для автоматической классификации документов государственных закупок по ОКПД2 кодам.
6. Проведение экспериментального исследования, включающего тестирование программного обеспечения на реальных данных о государственных закупках и сравнение результатов классификации с результатами, полученными при использовании других методов классификации.
7. Выводы и рекомендации по дальнейшему развитию и применению разработанного алгоритма классификации в различных областях, связанных с государственными закупками.

Методы

Некоторые методы, которые могут быть применены в рамках данной статьи, были описаны в моем предыдущей главе. Здесь я предоставлю более конкретный список методов и источников, которые могут быть полезны при работе над этой темой:

- Методы машинного обучения для классификации текстовых данных:
- [1] [2] [3]
- Методы предобработки текстовых данных: [4] [5]
- Методы визуализации данных: [6]
- Методы обработки естественного языка: [7]
- Методы оптимизации гиперпараметров: [8] [9]
- Методы метрической оценки: [10] [11]
- Методы при решении задач в области государственных закупок:
- [12] [13] [14]

Описание доменной области

ОКПД2 код – структура, показывающая принадлежность договора к определённому типу товара или услуги. Примеры целевых классов представлены таблице 1:

Таблица 1 – Структура ОКПД2 кода

Глубина кодировки	Тип	Пример кода	Значение кода
XX	Класс	14	одежда
XX.X	Подкласс	14.1	одежда, кроме одежды из меха
XX.XX	Группа	14.13	одежда верхняя прочая
XX.XX.X	Подгруппа	14.13.3	одежда верхняя прочая женская или для девочек
XX.XX.XX	Вид	14.13.34	платья, юбки и юбки-брюки женские или для девочек из текстильных материалов, кроме трикотажных или вязаных
XX.XX.XX.XX	Категория		
XX.XX.XX.XXX	Подкатегория	14.13.34.110	платья женские или для девочек из текстильных материалов, кроме трикотажных или вязаных

Конвейер сервиса

Сбор ссылок на документы

Для сбора сырых данных был определен источник данных - в данном случае это была веб-страница с необходимыми данными, размещенная на портале [15] с собранными в одном месте json со ссылками на документы и ОКПД2 кодами.

Для веб-страницы был создан отдельный запрос, который позволял получить нужные данные из HTML-кода страницы. Для этого была использована библиотека BeautifulSoup, которая позволяет парсить HTML-код и извлекать нужные элементы, в нашем случае это ссылки на архивы с JSON. Также была использована библиотека Requests, которая позволяет отправлять HTTP-запросы на сервер и получать ответы.

После того, как была собрана информация с веб-страницы, ссылки на документы и ОКПД2-коды были преобразованы и сохранены в файл формата CSV. Данный формат был выбран из-за его удобства (есть возможность добавлять записи в параллельном режиме) для дальнейшей обработки и анализа.

Таким образом, метод сбора сырых данных на основе сбора веб-страниц с использованием Python-библиотек позволяет собирать большое количество данных за короткий промежуток времени и обеспечивает удобство для дальнейшей обработки и анализа данных.

Сбор данных документов

Из исходного CSV извлекаются ссылки на документы, затем определяется формат документа "на лету". В данном процессе существует проблема того, что документы могут являться архивами, что затрудняет извлечения текста из этих файлов. Поэтому для упрощения на данном этапе принято использовать только doc, docx, pdf форматы. Файлы выгружаются в отдельную папку.

Сбор данных из документов

После сбора документов в форматах doc, docx, pdf - необходимо получить текст из них. На практике при сборе данных могут возникнуть трудность, когда текст не представлен в виде символов, а изображениями. Для решения данной проблемы достаточно простым, удобным, быстрым в реализации инструментом является textract. В этой библиотеке можно выбрать OCR-библиотека tesseract, она достаточно точно решает эту проблему.

Предобработка данных

Прежде чем приступить к классификации документов, следует выполнить подготовительную работу с данными. Целью предобработки является избавление семплов от выбросов и нормализация текста.

Обучение модели

Для обучения модели используется fasttext. Этот алгоритм достаточно нетребователен к ресурсам и выполняется достаточно быстро, а так же хорошо масштабируется.

Архитектура приложения

Описав выше все программы, присутствующие в данном приложении – построим схему его архитектуры на рисунке 1.

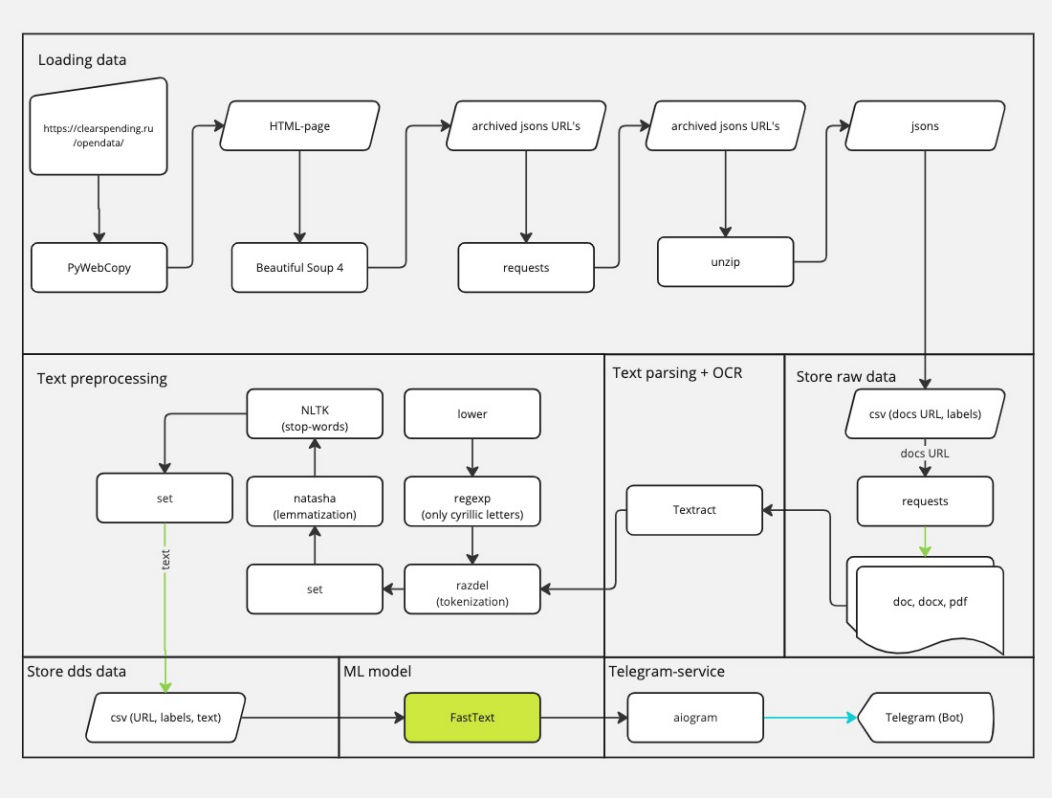


Рисунок 1 – Архитектура приложения

Скриншоты экранных форм

Приведём пример использования чат-бота в Telegram со вводом текста в сообщении на рисунке 2. Так же доступна подача данных документами.

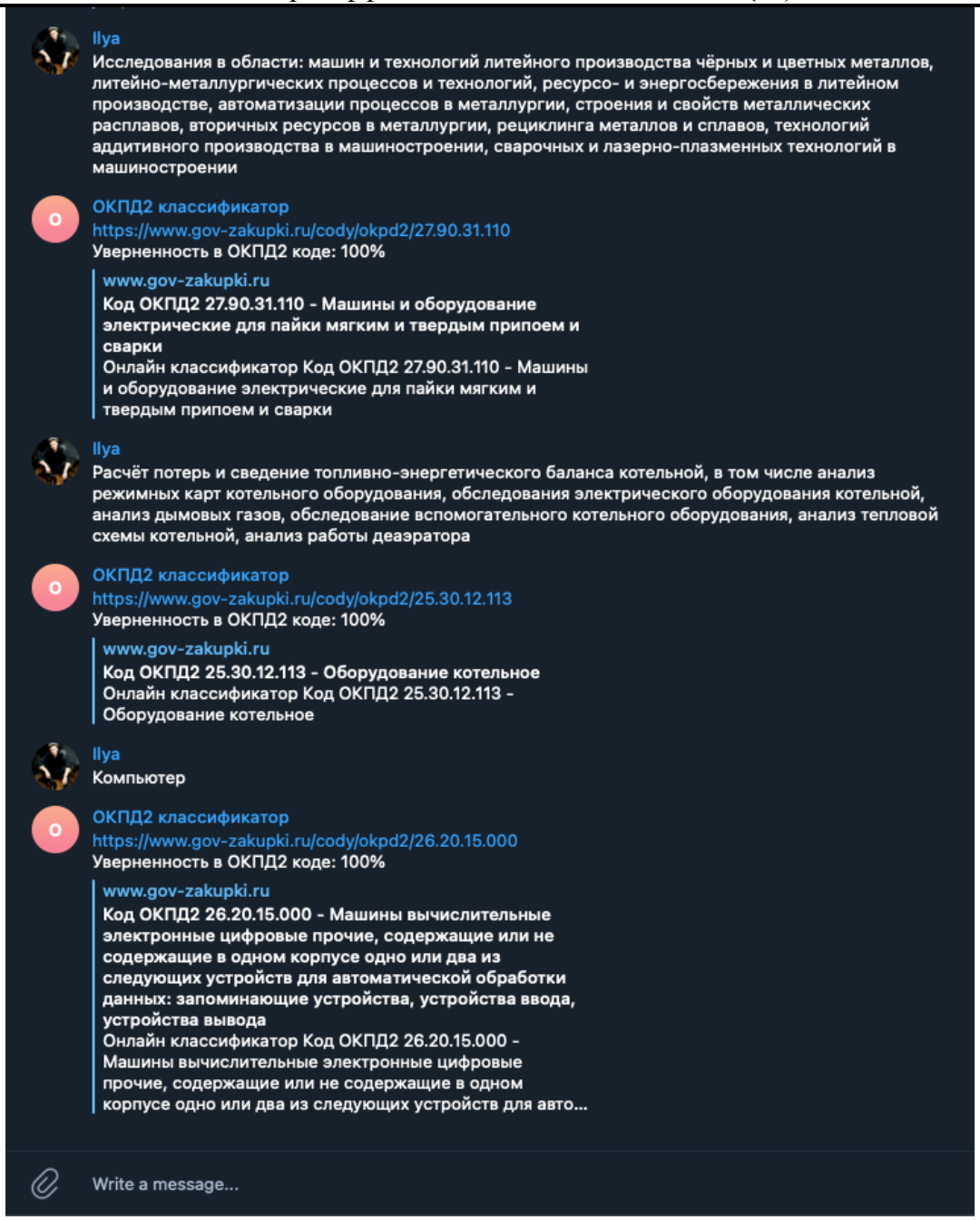


Рисунок 2 – Скриншот экранной формы

Результаты опытных испытаний

При сборе ссылок на документы получено порядка 30гб сырых json файлов. Далее эти файлы были объединены в единый csv файл с целевыми атрибутами (URL документа, ОКПД2 код).

Затем проведен анализ целевых классов, где установлен их сильный дисбаланс. Поэтом порядка из 9000 лейблов с количеством более 500 ссылок на документы случайным образом отобраны семплы до этого количества, остальные остаются как были. Данное число подбиралось для получения порядка 1'000'000 семплов при исходном порядке 10'000'000.

После получения данных для загрузки запущен процесс выгрузки документов. Даже при малом количестве семплов на класс получилось порядка 1 терабайта сырых данных

документов с учетом фильтрации документов по формату. Загрузка производилась в параллельном режиме в 20 поточном режиме порядка 2х недель.

Затем для загруженных файлов запущен процесс по получению текста из файлов и предобработке документов, который длился около недели в четырёхпоточном режиме. Текст добавлялся в csv файл со ссылками на документы и лейблами. В результате получено порядка 6 Гб обработанного csv файла.

Далее проведен анализ csv файла. Сначала исключаются дубликаты семплов, а затем проводится анализ количества токенов. При анализе обнаружилось, что в некоторой доле файлов достаточно мало токенов. Поэтому часть токенов удалено из распределения графическим способом. А т.е. Дано исходное распределение на рисунке 3А. Далее на рисунке 3Б исключены большие документы (более 13000 токенов),а затем на рисунке 3В исключены документы с менее 4800 токенов.

Теперь данные готовы для подготовки к обучению. Для этого разделим выборку в соотношении тренировочной и проверочной выборки в соотношении 95% к 5%. А новую тренировочную в соотношении 90% к 10%.

Предпоследним шагом при подготовке модели является обучение. При обучении на различных линейных, деревянных и прочих простых моделях возникали проблемы с ресурсами. Обучение производилось порядка 12 дней, с помощью FastText. При применении большинства алгоритмов возникали проблемы с ОЗУ, что показывает его экономичность по сравнению с другими алгоритмами.

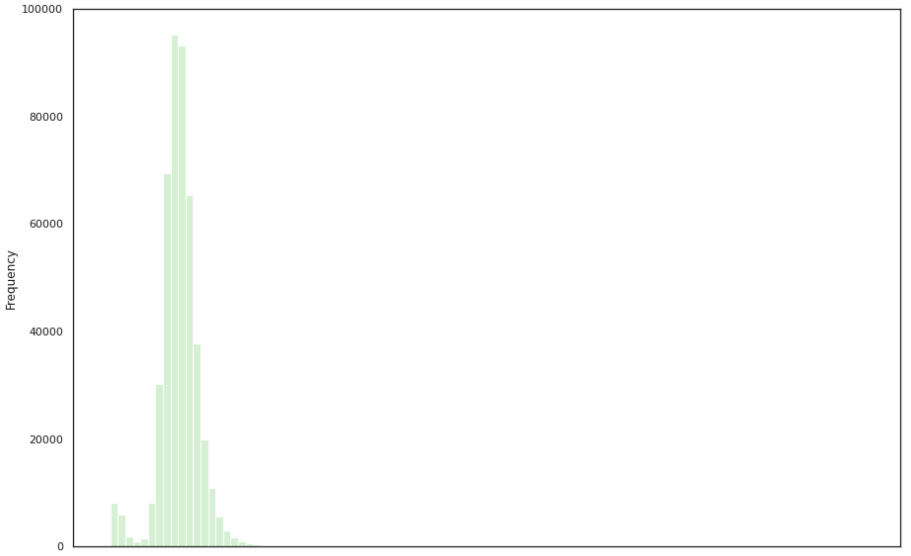
Результаты практического использования создаваемого продукта

На текущий момент прорабатываются возможные сенарии для применения данной модели. Потенциально данная модель может дополнить или заменить сервисы для поиска ОКПД2 коду, по ключевым словам, и прочему описанию. Так же в архитектуру можно добавить REST API и передавать предсказания пользователям по этому каналу передачи данных.

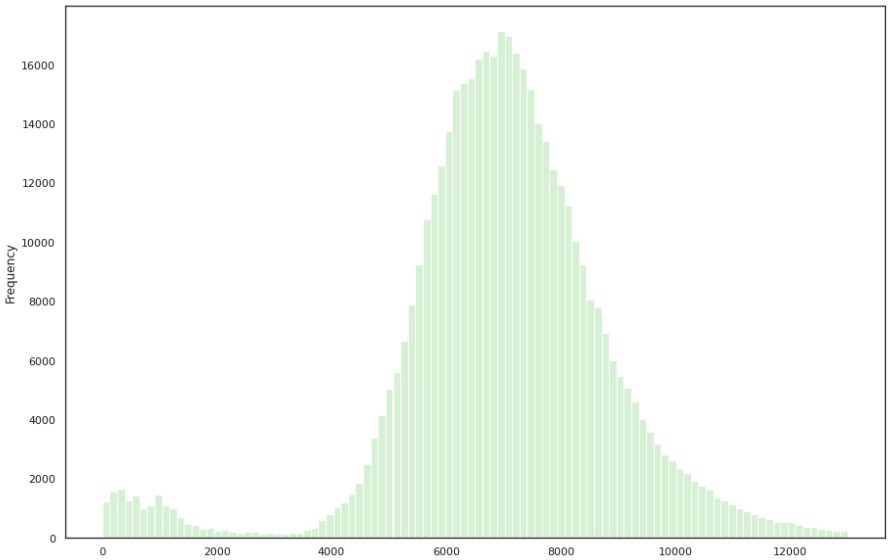
Таблицы и графики, подтверждающие работоспособность и достижение заданных целей

В конечном итоге получена метрики для полноценных кодов (2492 шт.) F1 - score равная 0.48, precision 0.48, recall 0.48.

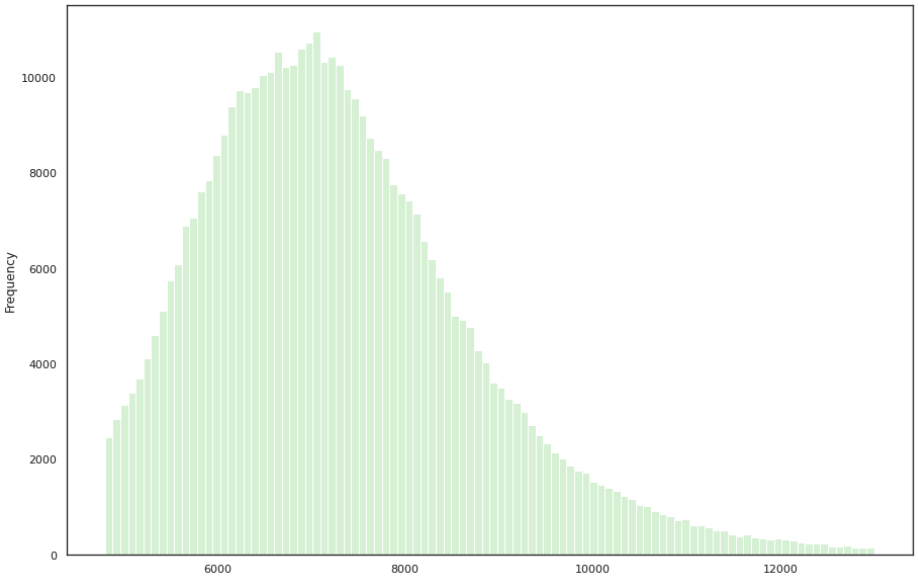
В таблице 2 показаны метрики первого элемента кода, а на рисунке 5 матрица согласия.



А)



Б)



В)

Рисунок 3 – Распределение количества токенов в каждом семпле

Так же построено облако слов для изучения семантического ядра (рисунок 4).

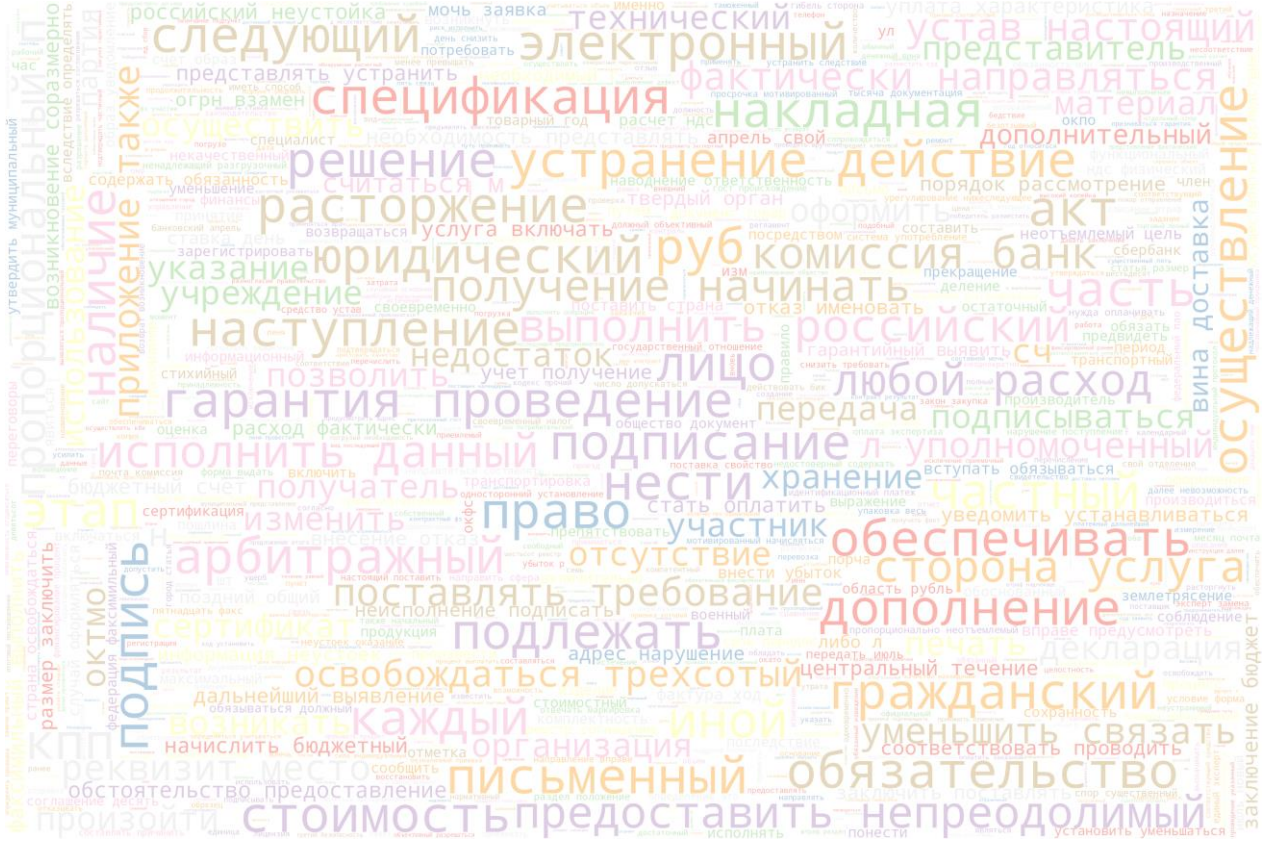


Рисунок 4 – Визуализация семантического ядра данных

Направления дальнейших исследований

- Применение более сложных и точных моделей.
- Погружение в доменную область и декомпозиция кодов.
- Исследование и применение других алгоритмов предобработки документов.
- Запуск периодических процессов загрузки документов в т.ч. подход machine unlearning для исключения устаревших кодов.
- Расширение числа поддерживаем форматов для распознавания.
- Увеличение скорости загрузки документов.
- Увеличение скорости распознавания документов.
- Добавление REST API или брокера сообщений в архитектуру.

Таблица – 2 Метрики первого элемента кода

labels	precision	recall	f1-score	support
1	0.87	0.87	0.87	733
2	0.96	0.85	0.9	128
3	0.71	0.74	0.73	81
5	0.99	0.95	0.97	83
6	0.67	0.67	0.67	3
8	0.73	0.75	0.74	178
10	0.96	0.96	0.96	3951
11	0.78	0.76	0.77	59
13	0.79	0.8	0.79	730
14	0.84	0.85	0.85	631
15	0.86	0.79	0.82	265
16	0.67	0.66	0.66	175
17	0.83	0.79	0.81	747
18	0.77	0.75	0.76	147
19	0.92	0.91	0.92	503
20	0.77	0.78	0.77	1426
21	0.92	0.94	0.93	2045
22	0.67	0.68	0.67	868
23	0.75	0.69	0.72	519
24	0.71	0.72	0.71	239
25	0.7	0.71	0.71	953
26	0.85	0.86	0.85	1938
27	0.79	0.79	0.79	1380
28	0.81	0.77	0.79	1671
29	0.92	0.91	0.91	86
accuracy				
weighted avg		0.84		19539

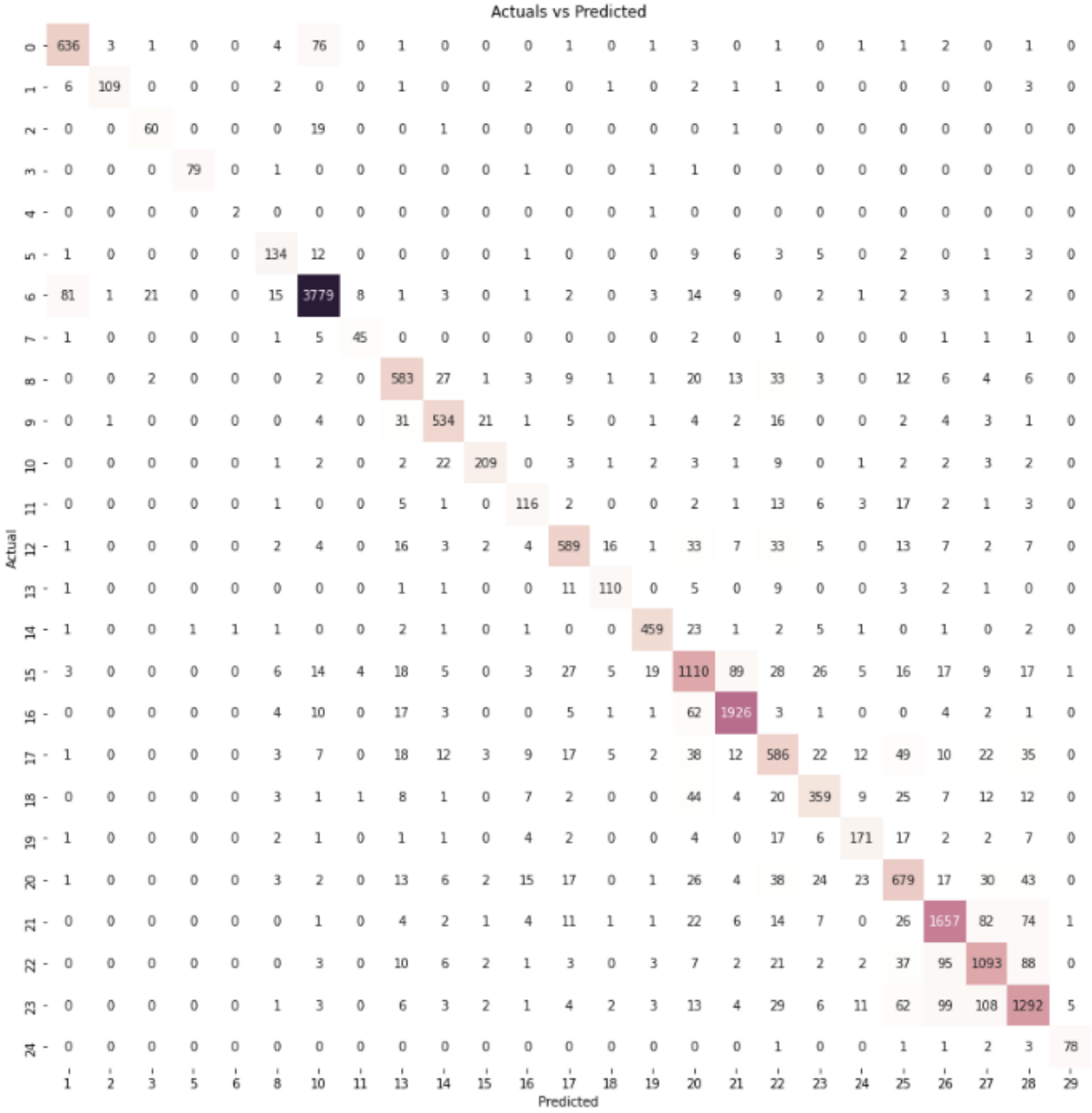


Рисунок 5 – Матрица согласия первого элемента кода

Заключение

Прежде всего, на основе анализа документов государственных закупок исследовано применение различных алгоритмов классификации документов. Выявлены программные и аппаратные ограничения большинства алгоритмов. Полученные в предыдущей главе метрики классификации показывают перспективность последующего исследования данной темы.

Список литературы

1. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval.
2. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. 3rd ed. Elsevier Inc., 2012.
3. Domingos P. A Few Useful Things to Know about Machine Learning // Communications of the ACM, Vol. 55(10), Октябрь 2012. pp. 78–87.

4. Jurafsky , Martin H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Vol II. Upper Saddle River, (New Jersey). 2008.
5. Manning D., Schutze H. Foundations of Statistical Natural Language Processing. II ed. Лондон: The MIT Press, 2000.
6. Tufte R. THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION. 3rd ed. Graphics Press, 2001. pp. 1-197
7. Eisenstein J. Introduction to Natural Language Processing. Cambridge: MIT Press, 2019. 1-536 pp.
8. Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research. 2012. Vol. 13. pp. 281-305.
9. Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arxiv.org. 2017. URL: https://arxiv.org/abs/1704.04861 (дата обращения: 1.Февраль.2023).
10. Gelman A., Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models // Cambridge University Press, 2007. pp. 1-625.
11. Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. Кембридж: Cambridge University Press, 2004 г. pp.1-462
12. Межов М.С. Использование машинного обучения для определения контировок, исходя из экономического смысла закупочной документации // International Journal of Open Information Technologies, Vol. 10, 2022. pp. 86-91.
13. Елисеев Д., Романов Д. Машинное обучение: прогнозирование рисков госзакупок // Открытые системы. Апрель 2018. Vol. 2.
14. Классификация данных в системе управления закупками // StecPoint. URL: https://stecpoint.ru/Practices-Classification/ (дата обращения: 28.Январь.2023).
15. Открытые данные [Электронный ресурс] // ГосЗатраты: [сайт]. [2023]. URL: https://clearspending.ru/opendata/ (дата обращения: 02.03.2023).

References

1. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval.
2. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. 3rd ed. Elsevier Inc., 2012.
3. Domingos P. A Few Useful Things to Know about Machine Learning // Communications of the ACM, Vol. 55(10), Oktyabr' 2012. pp. 78–87.
4. Jurafsky , Martin H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Vol II. Upper Saddle River, (New Jersey). 2008.
5. Manning D., Schutze H. Foundations of Statistical Natural Language Processing. II ed. London: The MIT Press, 2000.
6. Tufte R. THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION. 3rd ed. Graphics Press, 2001. pp 1-197.
7. Eisenstein J. Introduction to Natural Language Processing. Cambridge: MIT Press, 2019. pp.1-536

8. Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research. 2012. Vol. 13. pp. 281-305.
9. Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arxiv.org. 2017. URL: https://arxiv.org/abs/1704.04861 (data obrashcheniya: 1.Fevral'.2023).
10. Gelman A., Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models // Cambridge University Press, 2007. pp. 1-625.
11. Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. Kembridzh: Cambridge University Press, 2004 g. pp.1-462
12. Mezhov M.S. Ispol'zovanie mashinnogo obucheniya dlya opredeleniya kontirovok, iskhodya iz ekonomicheskogo smysla zakupochnoj dokumentacii // International Journal of Open Information Technologies, Vol. 10, 2022. pp. 86-91.
13. Eliseev D., Romanov D. Mashinnoe obuchenie: prognozirovanie riskov goszakupok // Otkrytye sistemy. Aprel' 2018. Vol. 2.
14. Klassifikaciya dannyh v sisteme upravleniya zakupkami // StecPoint. URL: https://stecpoint.ru/Practices-Classification/ (data obrashcheniya: 28.YAnvar'.2023).
15. Otkrytye dannye [Elektronnyj resurs] // GosZatraty: [sajt]. [2023]. URL: https://clearspending.ru/opendata/ (data obrashcheniya: 02.03.2023).