Институт проблем управления им. В.А. Трапезникова РАН
Лаборатория 42 “Интеллектуального анализа данных”

Интеллектуальный анализ больших объемов слабоструктурированных документов: модели, методы и перспективы исследований

Чехович Юрий Викторович,
к.ф.-м.н., с.н.с., заведующий лабораторией

26 февраля 2026 года

Структура доклада

1. Несколько слов об истории Лаборатории № 42
2. Об академических работах и задачах их анализа
3. Анализ структурных элементов работы
4. Поиск кросс-языковых совпадений и парафраза
5. Задачи детекции искусственного текста
6. Задачи поиска почти дубликатов изображений
7. Перспективные задачи анализа академических работ
8. Развитие теории сложности моделей и данных в моделях глубокого обучения

Несколько слов об истории Лаборатории № 42

- Лаборатория № 42 «Интеллектуального анализа данных» создана в ноябре 2025 года
- Коллектив Лаборатории (при создании) – «выпускники» научной школы академика Ю.И. Журавлева и академика К.В. Рудакова
- Основная тематика исследований научной школы – алгебраическая теория алгоритмов распознавания образов, классификации, прогнозирования, анализа данных, информационного поиска
- Команда лаборатории с 2005 по 2025 год занималась, в том числе, созданием и развитием первой отечественной системы обнаружения заимствований «Антиплагиат»

Об академических документах и задачах их анализа

Об академических работах

- ***Академические работы*** — это жанр письменных текстов, которые создаются для представления результатов научных исследований, анализа или обзора по конкретной теме, а также демонстрация компетенции автора(ов) в определенной области знаний в соответствии с научными и образовательными стандартами
- Включают **научные работы** и **учебные работы высшего профессионального образования** (курсовые работы, ВКР)
- Ежегодно в мире создается **несколько миллионов научных работ** (по оценкам до 5 млн в 2025 году и несколько десятков миллионов учебных работ
- Около 100 тысяч статей ежегодно публикуется в научных журналах с высокими наукометрическими показателями
- В базах данных содержится информация о порядка 100 млн научных публикаций высокого качества

Типы академических работ

Научные работы

Исследовательская статья

Обзорная статья

Заметка

Тезисы доклада

Полный текст доклада

Кандидатская диссертация

Докторская диссертация

Монография

Заявка на грант

Отчет по НИР/НИОКР

…

Учебные работы

Курсовая работа

Дипломная работа

Магистерская диссертация

…

Учебно-методические работы

Учебное пособие

Учебник

Методические рекомендации

…

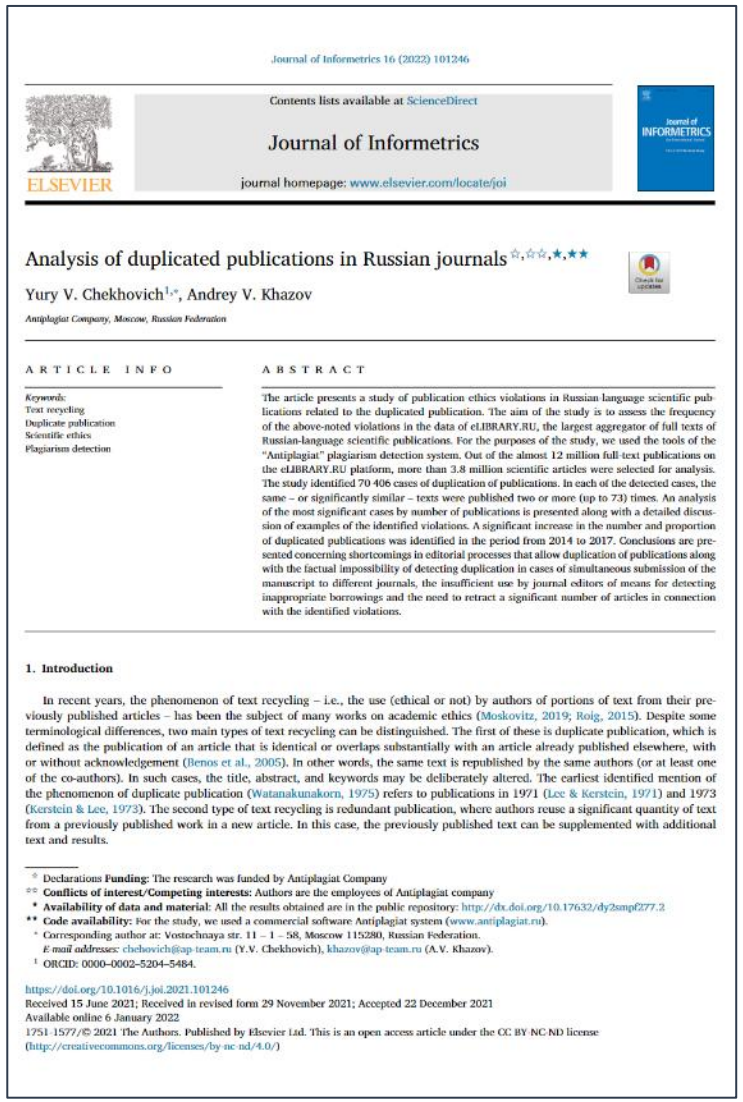
Состав и структура академической работы

Текст

Данные

Формулы

Изображения



Код программ

Ссылки

Цитаты

Структура

Journal of Informetrics 16 (2022) 101246

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Analysis of duplicated publications in Russian journals ☆,☆☆,★,★★

Yury V. Chekhovich[1,*], Andrey V. Khazov

Antiplagiat Company, Moscow, Russian Federation

ARTICLE INFO

Keywords:
Text recycling
Duplicate publication
Scientific ethics
Plagiarism detection

ABSTRACT

The article presents a study of publication ethics violations in Russian-language scientific publications related to the duplicated publication. The aim of the study is to assess the frequency of the above-noted violations in the data of eLIBRARY.RU, the largest aggregator of full texts of Russian-language scientific publications. For the purposes of the study, we used the tools of the "Antiplagiat" plagiarism detection system. Out of the almost 12 million full-text publications on the eLIBRARY.RU platform, more than 3.8 million scientific articles were selected for analysis. The study identified 70 406 cases of duplication of publications. In each of the detected cases, the same – or significantly similar – texts were published two or more (up to 73) times. An analysis of the most significant cases by number of publications is presented along with a detailed discussion of examples of the identified violations. A significant increase in the number and proportion of duplicated publications was identified in the period from 2014 to 2017. Conclusions are presented concerning shortcomings in editorial processes that allow duplication of publications along with the factual impossibility of detecting duplication in cases of simultaneous submission of the manuscript to different journals, the insufficient use by journal editors of means for detecting inappropriate borrowings and the need to retract a significant number of articles in connection with the identified violations.

1. Introduction

In recent years, the phenomenon of text recycling – i.e., the use (ethical or not) by authors of portions of text from their previously published articles – has been the subject of many works on academic ethics (Moskovitz, 2019; Roig, 2015). Despite some terminological differences, two main types of text recycling can be distinguished. The first of these is duplicate publication, which is defined as the publication of an article that is identical or overlaps substantially with an article already published elsewhere, with or without acknowledgement (Benos et al., 2005). In other words, the same text is republished by the same authors (or at least one of the co-authors). In such cases, the title, abstract, and keywords may be deliberately altered. The earliest identified mention of the phenomenon of duplicate publication (Watanakunakorn, 1975) refers to publications in 1971 (Lee & Kerstein, 1971) and 1973 (Kerstein & Lee, 1973). The second type of text recycling is redundant publication, where authors reuse a significant quantity of text from a previously published work in a new article. In this case, the previously published text can be supplemented with additional text and results.

☆ Declarations **Funding:** The research was funded by Antiplagiat Company
☆☆ **Conflicts of interest/Competing interests:** Authors are the employees of Antiplagiat company
★ **Availability of data and material:** All the results obtained are in the public repository: http://dx.doi.org/10.17632/dy2smpf277.2
★★ **Code availability:** For the study, we used a commercial software Antiplagiat system (www.antiplagiat.ru).
* Corresponding author at: Vostochnaya str. 11 – 1 – 58, Moscow 115280, Russian Federation.
E-mail addresses: chehovich@ap-team.ru (Y.V. Chekhovich), khazov@ap-team.ru (A.V. Khazov).
[1] ORCID: 0000-0002-5204-5484.

https://doi.org/10.1016/j.joi.2021.101246
Received 15 June 2021; Received in revised form 29 November 2021; Accepted 22 December 2021
Available online 6 January 2022

Системы поддержки экспертных решений

- Жизненный цикл каждой академические работы требует проведения экспертизы (оценка редактора, рецензирование, отзыв, заключение, оценка качества)
- Экспертиза включает в себя выявление возможных нарушений в работе: неправомерные заимствования (плагиат), некорректное цитирование, включая некорректное самоцитирование, фабрикацию данных, фальсификацию результатов, дублирование публикаций и т.п.
- Системы обнаружения заимствований призваны обеспечивать эксперта инструментарием и данными
- Фактически использование систем часто сводится к автоматизированной оценке доли текста, совпавшего с источниками, и сопоставлении этой с каким-либо порогом
- Специфика типа работы или информация о внутренней структуре работы не используется, кроме того, инструменты информационного поиска работают только с текстом и выявляют только «прямые совпадения» фрагментов документа с источниками

Типы систем обнаружения заимствований

External plagiarism detection

Поиск совпадений с внешними источниками

Требует обширной базы данных

Сравнение текста с миллиардами документов

Intrinsic plagiarism detection

Анализ стилистических особенностей текста

Поиск аномалий в рамках одного документа

Не требует внешней базы источников

Структура алгоритма поиска заимствований

Извлечение и предобработка текста

Удаление лишних символов, токенизация, удаление стоп-слов, лемматизация

Шинлирование

Формирование n-грамм (шинглов) из последовательных слов

Хеширование

Преобразование шинглов в числовые значения

Поиск кандидатов

Поиск по индексу, формирование списка документов-кандидатов

Редуцирование множества кандидатов

Разностный алгоритм

Сопоставление кандидатов с запросом

Формирование отчета

Анализ структурных элементов работы. Метаданные

Цель исследования — провести анализ текста научной работы с целью извлечения метаданных, которые могут быть использованы для проверки работы.

Категория	Метод	P	R	F
Заголовок	Предложенный	**0,74**	**0,79**	**0,76**
	Предложенный только на текстовых признаках	0,67	0,66	0,66
	Базовый	0,20	0,77	0,32
Автор	Предложенный	**0,78**	0,71	**0,74**
	Предложенный только на текстовых признаках	0,45	0,74	0,56
	Базовый	0,33	**0,75**	0,46

Качество модели извлечения метаданных из текста научной работы.
В качестве истинных меток использовалась асессорская разметка научных работ.

А. В. Огальцов, О. Ю. Бахтеев Автоматическое извлечение метаданных из научных PDF-документов // Информатика и ее применения. – 2018. – Т. 12, № 2. – С. 75-82. – DOI 10.14357/19922264180211. – EDN XROLVB

Анализ структурных элементов работы

Цель исследования — разработка метода деления документа на структурные элементы

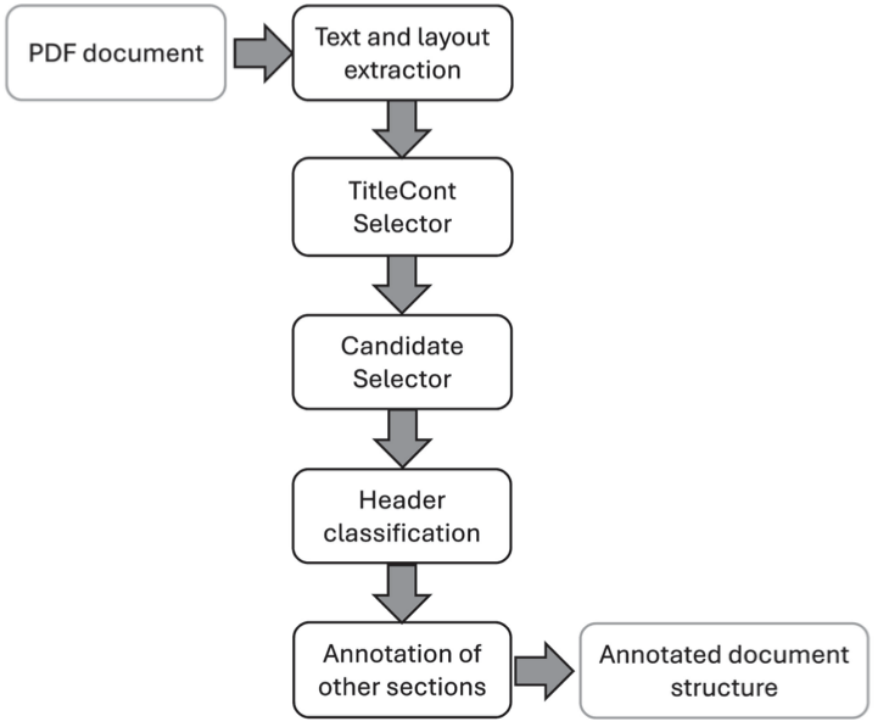


Схема извлечения структурных элементов в формате IMRAD

Section	Precision	Recall	F1
Title	0.97	0.88	0.92
Contents	0.99	0.84	0.91
Introduction	0.77	0.75	0.76
Methods	0.75	0.52	0.62
Results	0.81	0.26	0.40
Conclusion	0.85	0.82	0.83
Bibliography	0.88	0.95	0.91
Appendices	0.92	0.94	0.93

Качество нашей модели на мультиязычной выборке

Section	Precision	Recall	F1
Introduction	0.34	0.39	0.36
Methods	0.13	0.09	0.10
Results	0.08	0.08	0.08
Conclusion	0.09	0.11	0.10
Bibliography	0.79	0.32	0.46
Appendices	0.54	0.53	0.54

Качество базового метода извлечения GROBID

I. Kopanichuk, A. Chashchin, I. Ochneva, A. Grabovoy, A. Ogaltsov, A. Kildyakov, Y. Chekhovich "Structure Extractor: Multilingual Extraction of Sections from Scientific Document," 2025 37th Conference of Open Innovations Association (FRUCT), Narvik, Norway, 2025, pp. 122-128, doi: 10.23919/FRUCT65909.2025.11008070

Анализ структурных элементов работы. Библиография

Цель: построение метода извлечение метаданных из библиографических записей научных работ.

Границы фрагментов -> Сегментация на записи -> Разделение на поля

Извлекаемые поля:

1. Авторы — высокое качество извлечения.
2. Название — высокое качество извлечения.
3. Год — высокое качество извлечения.
4. Страницы — среднее качество извлечения.
5. Издатель — среднее качество извлечения.
6. Том журнала — низкое качество.
7. Выпуск журнала — низкое качество.

Поиск кросс-языковых заимствований и парафраза

Предпосылки

- Увеличение объемов доступной научной информации
- Развитие средств машинного перевода
- Развитие автоматизированных инструментов контроля за заимствованиями

Поиск кросс-языковых заимствований и парафраза

Цель исследований: построение методов поиска схожих, перефразированных и переводных текстовых фрагментов по текстовым коллекциям.

Задано: проверяемый документ и коллекция документов, с которыми происходит сравнения:

$$d = (d_1, \ldots, d_n), \qquad \mathfrak{C} = \{c_i\}_{i=1}^{L}, \quad c_i = (w_1^i, \ldots, w_{n_i}^i)$$

Требуется: предъявить отображения:

$$f : \mathfrak{D} \times \mathfrak{C} \to \mathfrak{S}, \quad \{(d_{s_k^d}, \ldots, d_{e_k^d}), (w_{s_k}^{c_{l_k}}, w_{e_k}^{c_{l_k}})\}_{k=1}^{K} \in \mathfrak{S}$$

Решение: декомпозиция задачи на части

1. Поиск документов кандидатов
2. Поиск фрагментов внутри документа и верификация

Р. В. Кузнецова, О. Ю. Бахтеев, Ю. В. Чехович Методы обнаружения переводных заимствований в больших текстовых коллекциях / // Информатика и ее применения. – 2021. – Т. 15, № 1. – С. 30-41. – DOI 10.14357/19922264210105. – EDN BQFZAZ.

Поиск кросс-языковых заимствований и парафраза

Преобразования текста в последовательность кластеров

1. Удаление стоп слов.
2. Замена слов на кластера синонимов, построенных по словарям и векторным выравниваниям

Построение шинглов (n-грам) кластеров

1. Выделение n-грамм слов.
2. Сортировка слов внутри шингла, для уменьшения влияния перестановки слов.
3. Построения хеш таблицы.

Алгоритм	Recall@10
Базовый	0,93
Представленный	0,95

Бахтеев О.Ю., Чехович Ю.В., Грабовой А.В., Горбачев Г.В., Горленко Т.А., Гращенков К.В., Ивахненко А.А., Кильдяков А.С., Хазов А.В., Комарницкий В.Е., Никитов А.В., Огальцов А.В., Сахарова А.В. Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works / Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Proceedings from the European Conference on Academic Integrity and Plagiarism 2021. Cham: Springer, 2022. С. 143-161.

Поиск кросс-языковых заимствований и парафраза

Получения векторных представлений фрагментов для поиска

1. Деление текста на “простые предложения” — минимальные смысловые фрагменты.
2. Векторизация каждого фрагмента при помощи модели обученной на триплет лоссе.

Попарное сравнение векторов

1. Расстояние между векторами определяется как косинусное расстояние.

Алгоритм	Precision	Recall	F1
Базовый	0,99	0,15	0,26
Представленный	0,93	0,80	0,85

- O. Bakhteev, R. Kuznetsova, A. Romanov and A. Khritankov, "A monolingual approach to detection of text reuse in Russian-English collection," 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), 2015, pp. 3-10
- Avetisyan, K., Gritsay, G. & Grabovoy, A. " Cross-Lingual Plagiarism Detection: Two Are Better Than One " Program Comput Soft 49, 346–354 (2023).
- Oleg Bakhteev Yury Chekhovich, Andrey Grabovoy et al. (2022). "Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works" In: Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Ethics and Integrity in Educational Contexts, vol 4. Springer, Cham.

Поиск заимствованных изображений

Предпосылки

- Существует множество научных работ, значительная часть содержания которых заключена в изображениях

Особенности

Распространенные изменения:

- Слабое кадрирование (до 5% от оригинального размера)
- Умеренное изменение пропорций (не более 25% от оригинального размера)
- Перевод в градации серого
- Размытие

Редкие изменения:

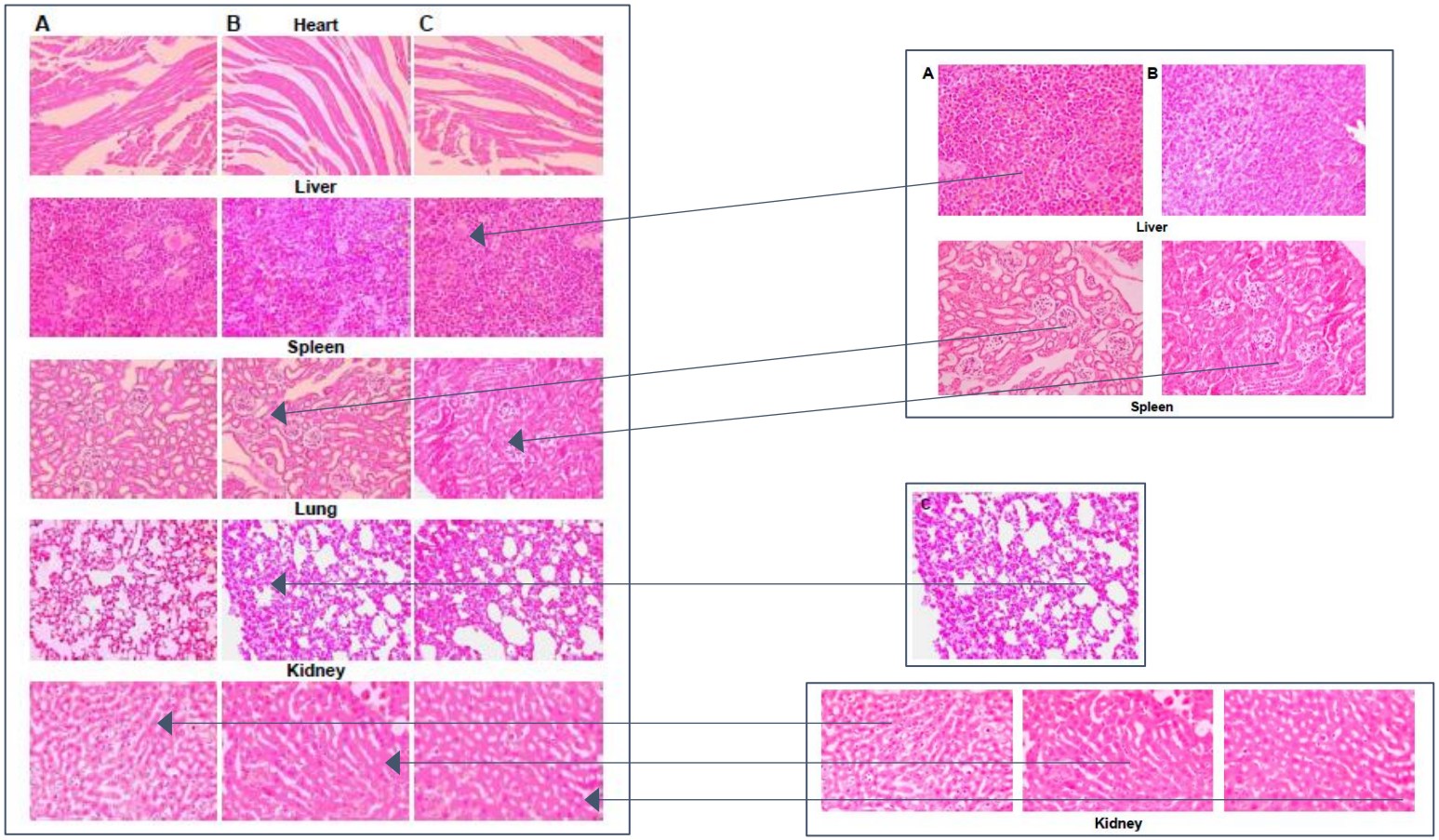
- Значительное кадрирование (до 15% от оригинального размера)
- Значительное изменение пропорций (не более 50% от оригинального размера)
- Поворот на 90, 180, 270 градусов

Поиск заимствованных изображений

- Построение векторного представления изображения.
  Использовалась архитектура Deep Ranking*
- Поиск кандидатов. Использовалась векторная база данных FAISS
- Точное сравнение кандидатов с запросом.
  Использовалась архитектура типа Siamese Neural Network

- М.С. Каприелова, К.В. Варламова, И.О. Потяшин, Ю.В. Чехович. Поиск почти дубликатов изображений рукописных текстов для высоконагруженных сервисов // Известия Российской академии наук. Теория и системы управления. - 2024. - №4. - С. 129-139.
- M. Kaprielova, A. Grabovoy, I. Potyashin, K. Varlamova, Y. Chekhovich. Image Plagiarism Detection Pipeline for Vast Databases // 2024 35th Conference of Open Innovations Association (FRUCT). – IEEE, 2024. – С. 328-335.

Поиск заимствованных изображений



Li S, Wang X. In vitro and in vivo evaluation of novel NGR-modified liposomes containing brucine. *Int J Nanomedicine*. 2017;12:5797-5804 https://doi.org/10.2147/IJN.S136378

Li D, Gong L. Preparation of novel pirfenidone microspheres for lung-targeted delivery: in vitro and in vivo study. *Drug Des Devel Ther*. 2016;10:2815-2821 https://doi.org/10.2147/DDDT.S113670

Chen J, Jiang H, Wu Y, Li Y, Gao Y. A novel glycyrrhetinic acid-modified oxaliplatin liposome for liver-targeting and in vitro/vivo evaluation. *Drug Des Devel Ther*. 2015;9:2265-2275 https://doi.org/10.2147/DDDT.S81722

Zhou X, Tao H, Shi KH. Development of a nanoliposomal formulation of erlotinib for lung cancer and in vitro/in vivo antitumoral evaluation. *Drug Des Devel Ther*. 2018;12:1-8 https://doi.org/10.2147/DDDT.S146925

Поиск заимствованных изображений

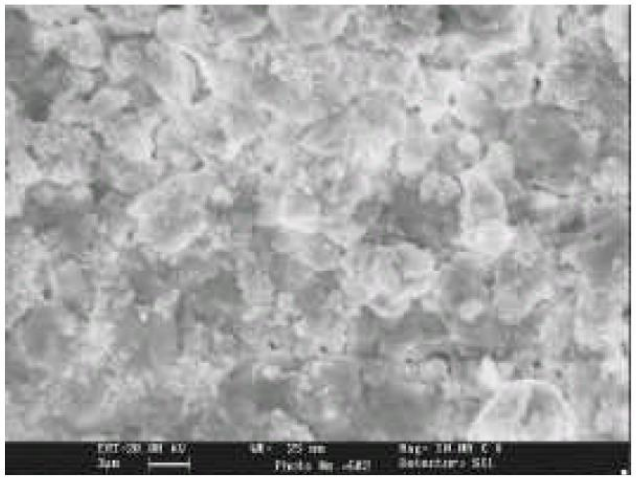


Fig. 5c. SEM image of MgSA6 composite.

J. JUDITH VIJAYA, L. JOHN KENNEDY, G. SEKARAN, K.S. NAGARAJA
Methanol Sensing Behavior of Strontium(II) Added MgAl2O4 Composites Through Solid-State Electrical Conductivity Measurements
Sensors & Transducers Journal, Vol.74, Issue 12, December 2006, pp.864-873

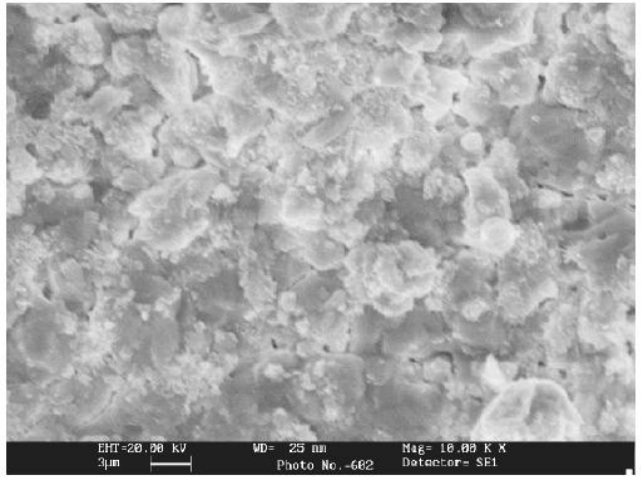
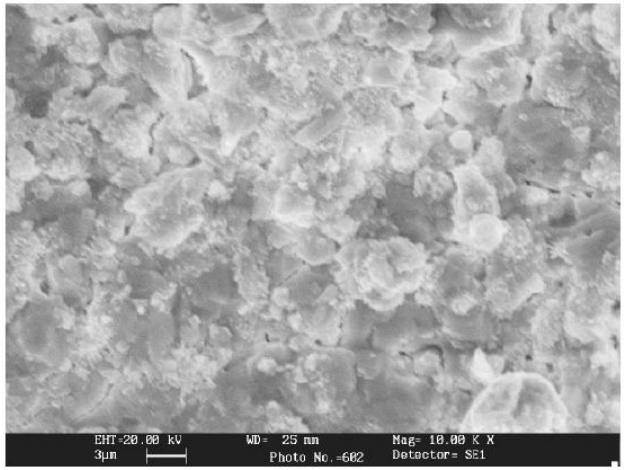


Fig. 5c. SEM image of ZnSA6 composite.

J. JUDITH VIJAYA, L. JOHN KENNEDY, G. SEKARAN, K.S. NAGARAJA
Synthesis, Characterization and Acetone Sensing Properties of Novel Strontium(II)-added ZnAl2O4 Composites Sensors & Transducers Journal, Vol.76, Issue 2, February 2007, pp.1008-1017

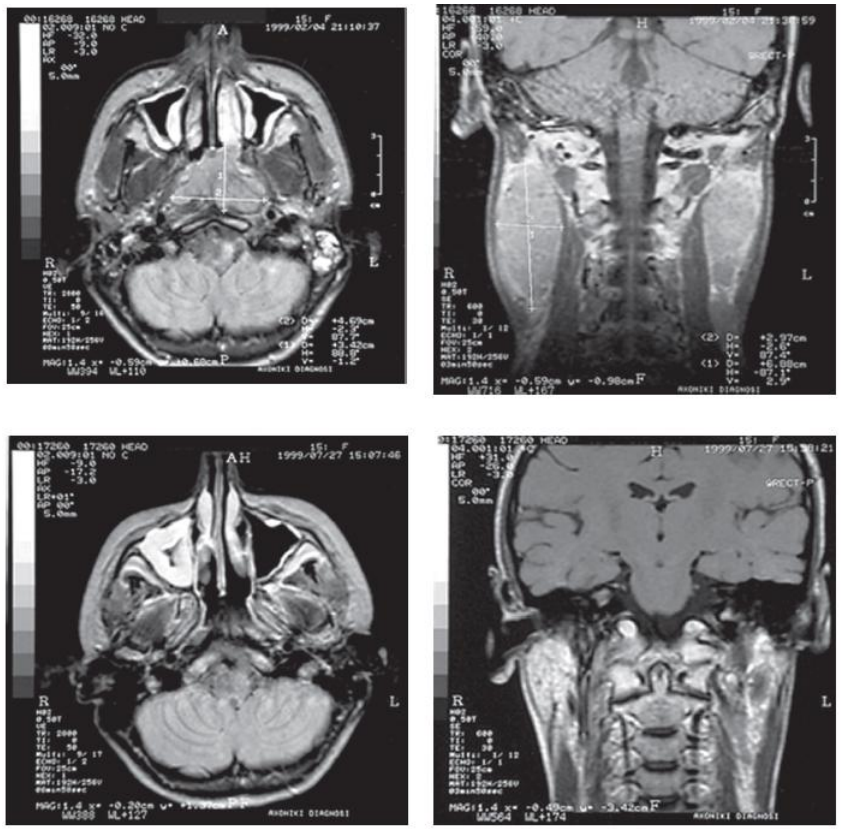


(c)

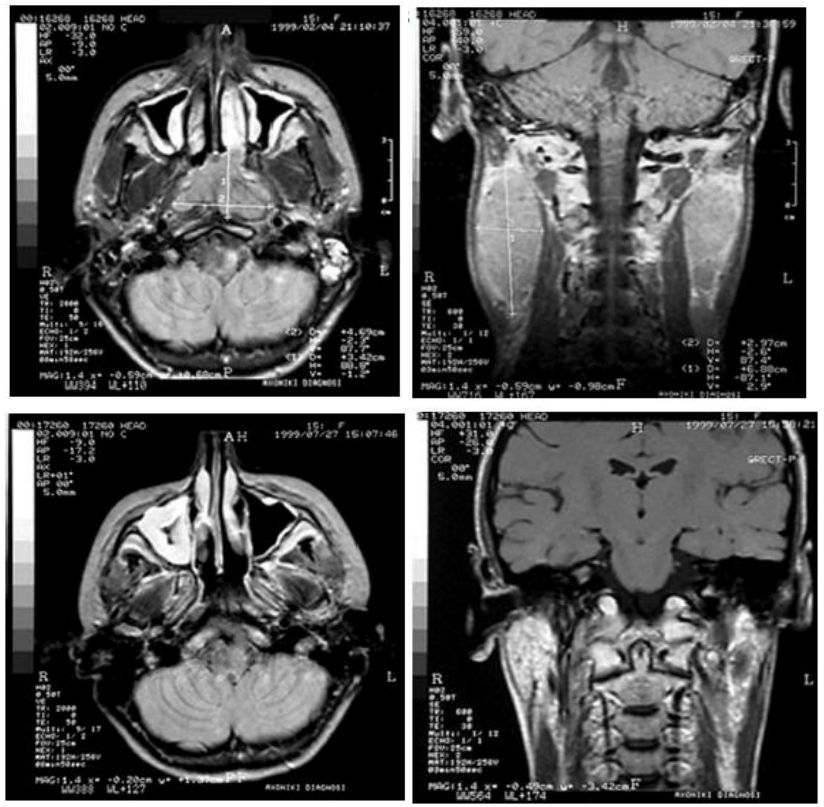
Fig. 3. SEM imag of a) BaSA1, b) BaSA5 and c) SA composite

B. Jeyaraj, L. John Kennedy, G. Sekaran and J. Judith Vijaya
Benzene and Toluene Vapor Sensing Properties of Sr(II)-added Barium Aluminate Spinel Composites
Sensors & Transducers Journal, Vol. 96, Issue 9, September 2008, pp. 68-80

Поиск заимствованных изображений

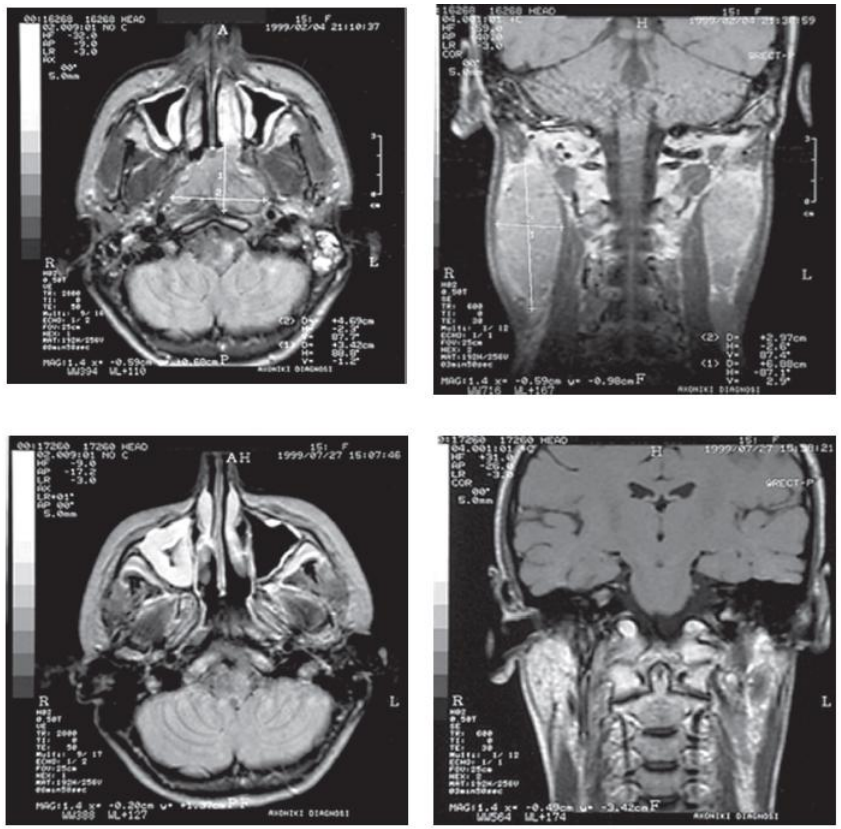


Nikolaos Eleftheriadis, Christos Papaloukas, Damianos Eleftheriadis, Apostolos Hatzitolios, Ioulia Ioannidou-Marathiotou & Kiki Pistevou-Gompaki Long-term radiotherapy related complications in children with head and neck cancer: Another era for pediatric oncologic pathology // International Journal of General Medicine 2009:2 63–66

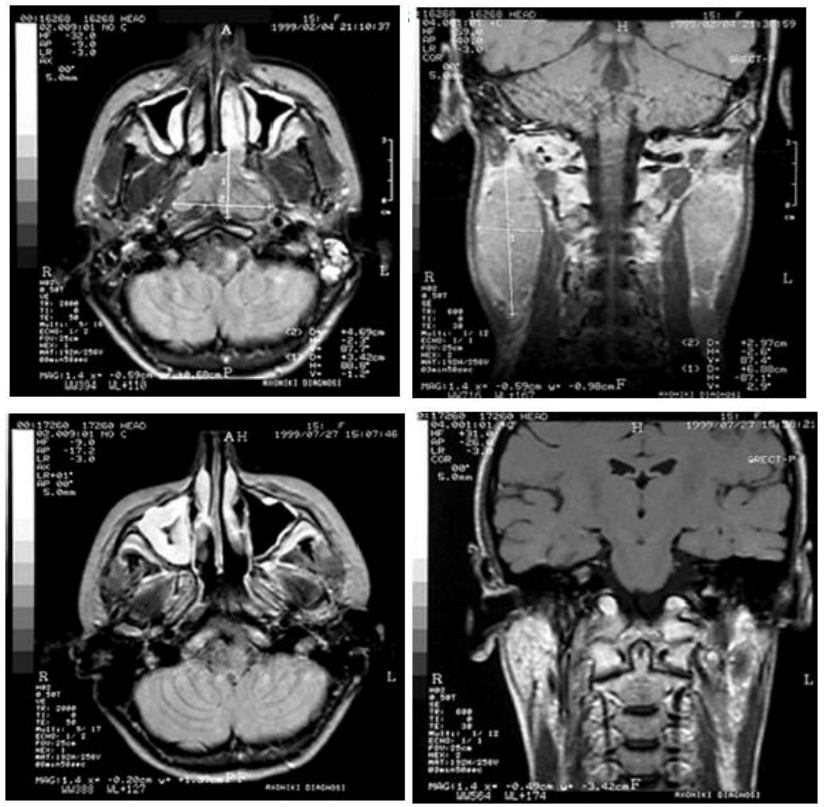


Ioulia Ioannidou-Marathiotou, Kyriaki Pistevou-Gompaki, Nikolaos Eleftheriadis & Christos Papaloukas Long term chemoradiotherapy-related dental and skeletal complications in a young female with nasopharyngeal carcinoma // International Journal of General Medicine 2010:3 187–196

Поиск заимствованных изображений



Abstract: ...We report on two male children (8 and 14 years old) with head and neck cancer, who were successfully treated with chemoradiotherapy and presented with growth deficiency of middle face and mandible hypoplasia, eight years and one year later, respectively...



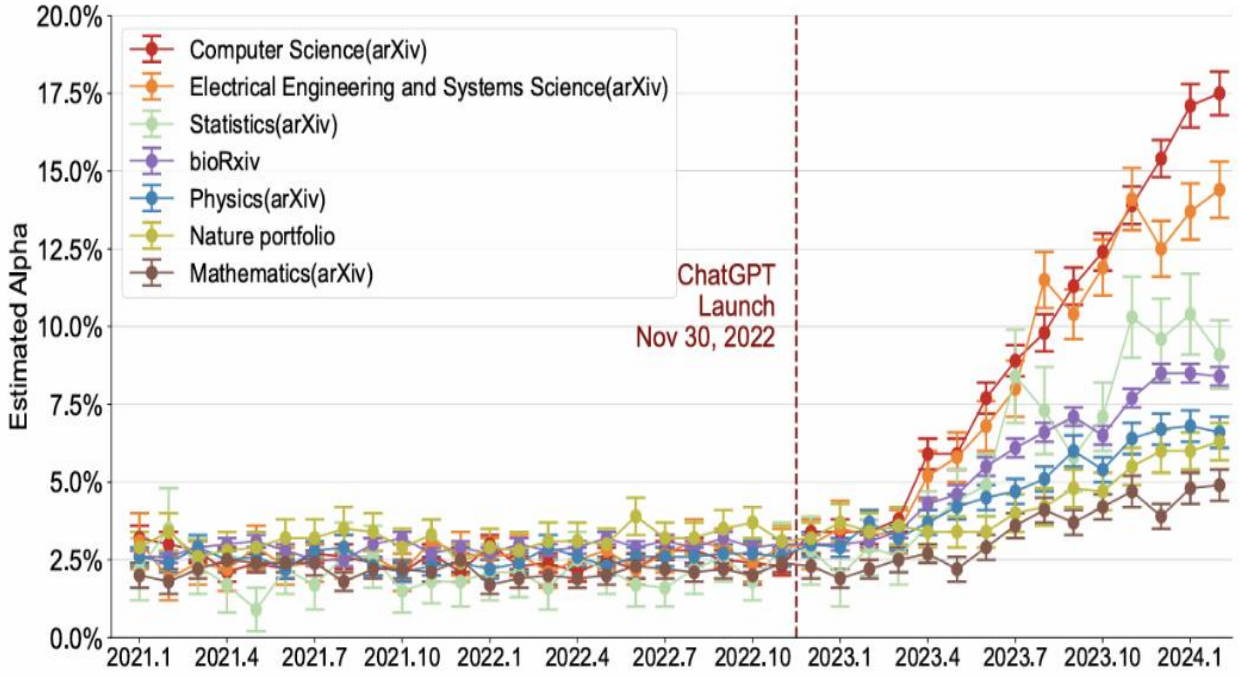
Abstract: We describe the long-term complications six years after chemoradiotherapy in a 20-year old woman with nasopharyngeal carcinoma. We wanted to know whether the radiation dose was constant throughout the oral cavity, and thus uniformly affecting the corresponding dental and skeletal structures...

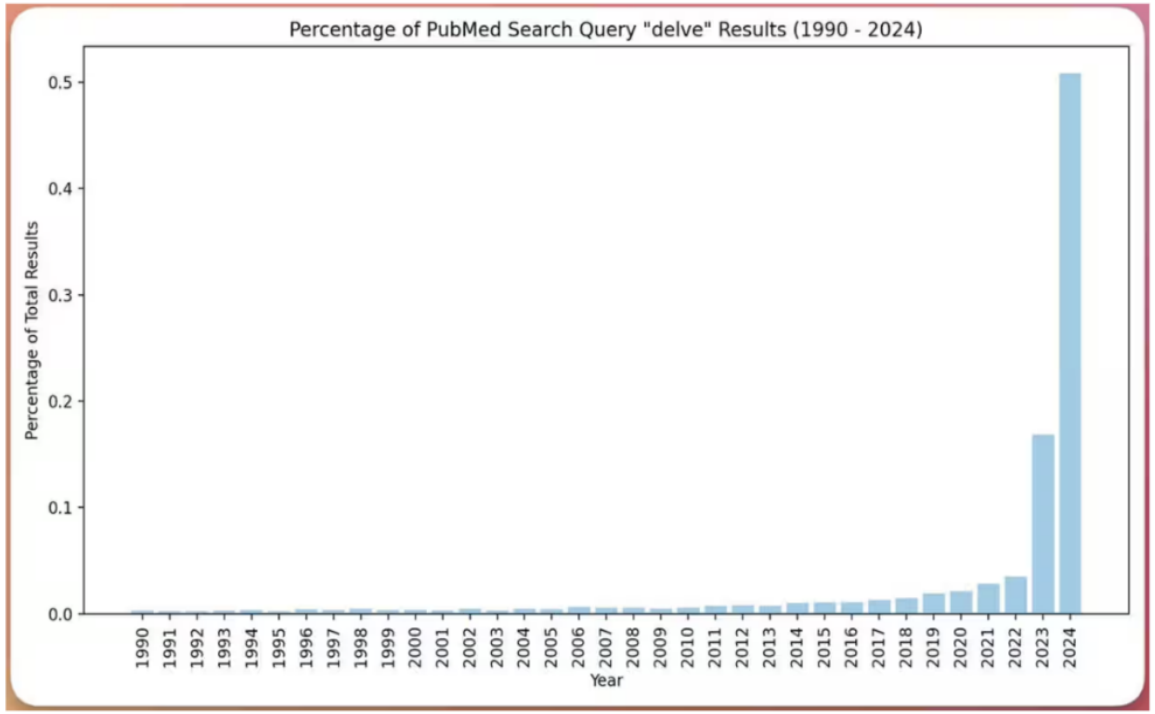
Детекция машинно сгенерированных текстов

Цель: построение методов поиска, верификации и интерпретации сгенерированных текстовых последовательностей.

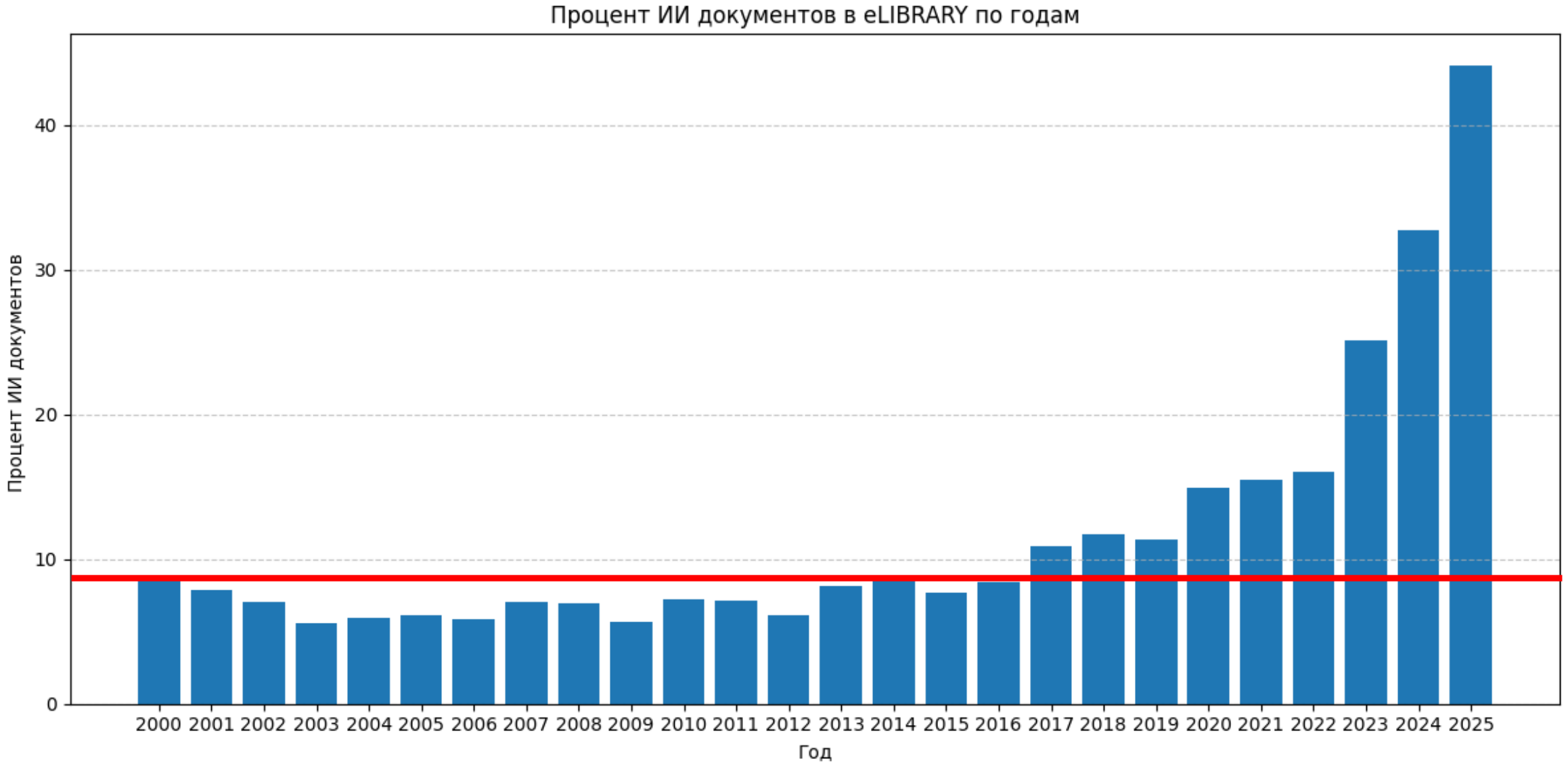
Требуется предложить: метод детектирования машинно-сгенерированных текстовых последовательностей, основанный на паттернах присущих искусственно созданным фрагментам, а также метод их интерпретации и верификации качества.

Актуальность темы





Детекция машинно сгенерированных текстов



Детекция машинно сгенерированных текстов

Развитие теории сложности моделей и данных в моделях глубокого обучения

Эволюция теории сложности обучения

Статистическая теория сложности (1970–2010)

▶ Вапник, Червоненкис, 1974 — введение VC-размерности как меры емкости класса; равномерная сходимость частот к вероятностям.
▶ Valiant, 1984 — формализация понятия обучаемости; гарантии с заданной точностью и надежностью.
▶ Воронцов, 2010 — развитие VC-теории; учет структуры данных и локализации алгоритмов для снижения консервативности оценок.

Аппроксимационная теория (1989–2015)

▶ Cybenko, 1989 — однослойная сеть с сигмоидой может аппроксимировать любую непрерывную функцию.
▶ Hastad, 1987; Bengio, 2007; Cohen, 2015– экспоненциальный рост выразительной способности с увеличением числа слоев.

Современные эмпирические подходы (2017–2022)

▶ Sagun, 2017; Keskar, 2016 — эмпирическое исследование спектра Гессе; связь “плоских” минимумов с обобщением.
▶ Kaplan, 2020; Hoffmann, 2022 — степенные зависимости качества от числа параметров и объема данных; оптимальные соотношения для вычислительных бюджетов.

Направления развивались независимо. Отсутствует единая теория, связывающая сложность модели и сложность данных.

Цель и задачи исследования

Цель: Построение единого теоретического аппарата для оценки сложности

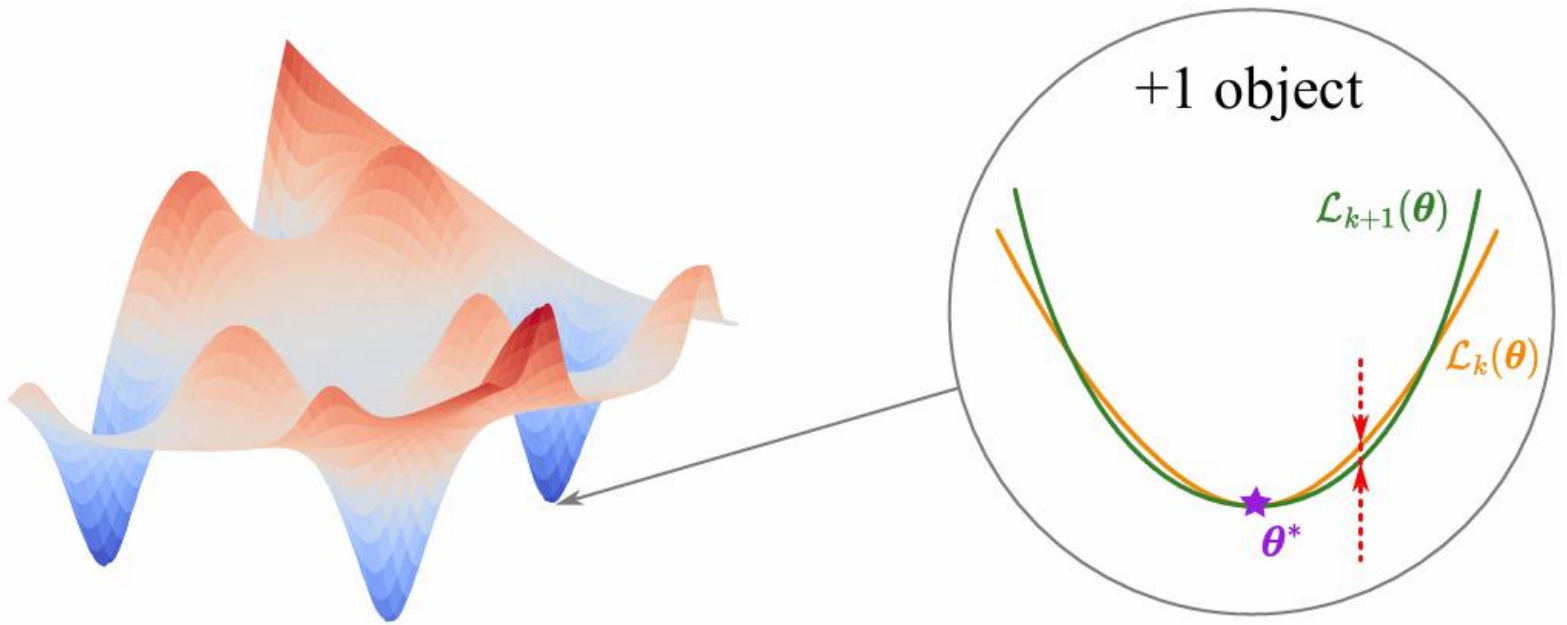
моделей глубокого обучения и сложности данных, а также установление формальных критериев соответствия между сложностью модели и сложностью

выборки, необходимой для ее обучения.

Задачи:

1. Введение формальных определений мер сложности моделей и данных в рамках теории мер и установление критерия обучаемости модели на выборке.
2. Получение теоретических оценок спектральных норм матриц Гессе для полносвязных, сверточных и трансформерных архитектур.
3. Построение ландшафтной меры сложности модели на основе анализа матриц Гессе и установление ее связи с условной сложностью выборки.
4. Построение методов оценки достаточного объема выборки на основе анализа стабильности функции потерь и близости апостериорных распределений параметров.
5. Построение методов снижения сложности моделей глубокого обучения на основе анализа матриц Гессе и методов дистилляции знаний.
6. Демонстрация практического применения построенного теоретического аппарата в прикладных задачах.

От общей теории к конкретной мере сложности



(a) Loss function landscape (b) Losses difference

Если добавление нового объекта данных существенно изменяет ландшафт оптимизации, то модель недостаточно обучена на текущей выборке.

Результаты

1. Единый теоретический аппарат оценки сложности моделей глубокого обучения и сложности данных на основе теории мер и анализа ландшафта оптимизационной задачи, включающий формальные определения мер сложности и критерий обучаемости модели на выборке.
2. Ландшафтная мера сложности модели через спектральные свойства матриц Гессе функции потерь.
3. Теоретические оценки ландшафтных мер на основе матриц Гессе для основных архитектур моделей глубокого обучения.
4. Оценки достаточного объема выборки и их связь со сложностью моделей.
5. Методы снижения сложности моделей глубокого обучения на основе анализа градиентов, дистилляции и анти-дистилляции для передачи знаний между моделями и доменами данных

Спасибо за внимание

Ваши вопросы

Чехович Юрий Викторович, к.ф.-м.н., с.н.с.,
заведующий лабораторией

26 февраля 2026 года