РЭУ.РФ
РОССИЙСКИЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Г.В. ПЛЕХАНОВА

Аналитический центр Москвы

Большие генеративные системы в официальной статистике: международный опыт использования и оценки с применением технологий text-mining

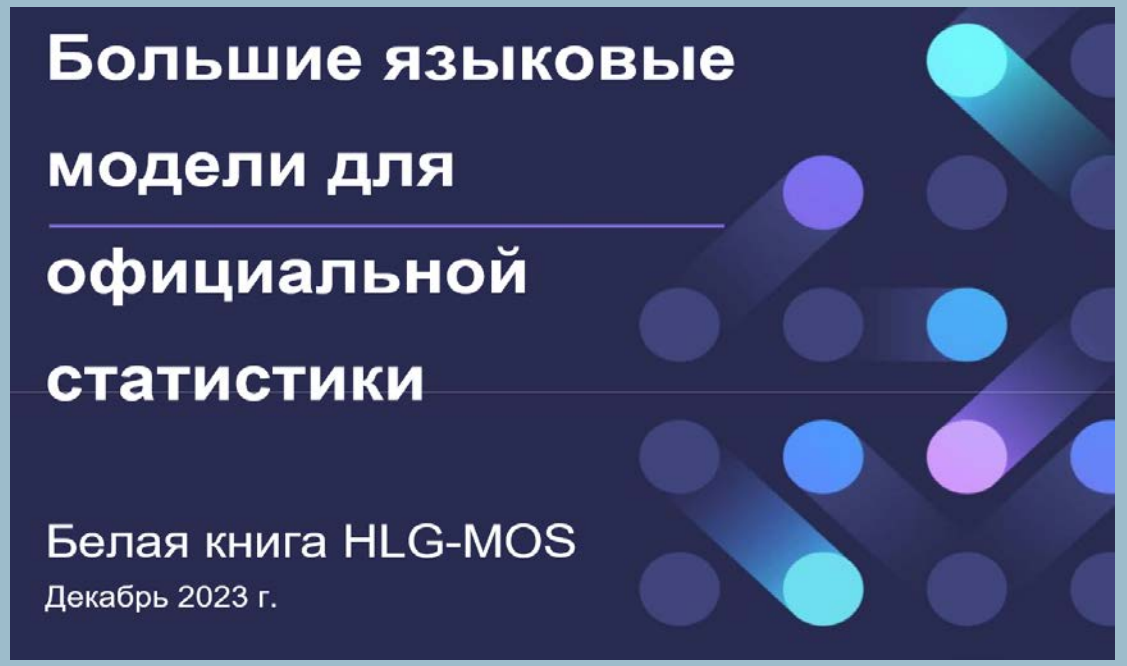
Зарова Е.В., д.э.н., профессор

Актуальность

2022-2023 гг. - «взрывная» волна интереса к Chart GPT, в том числе в официальной статистике

UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS)

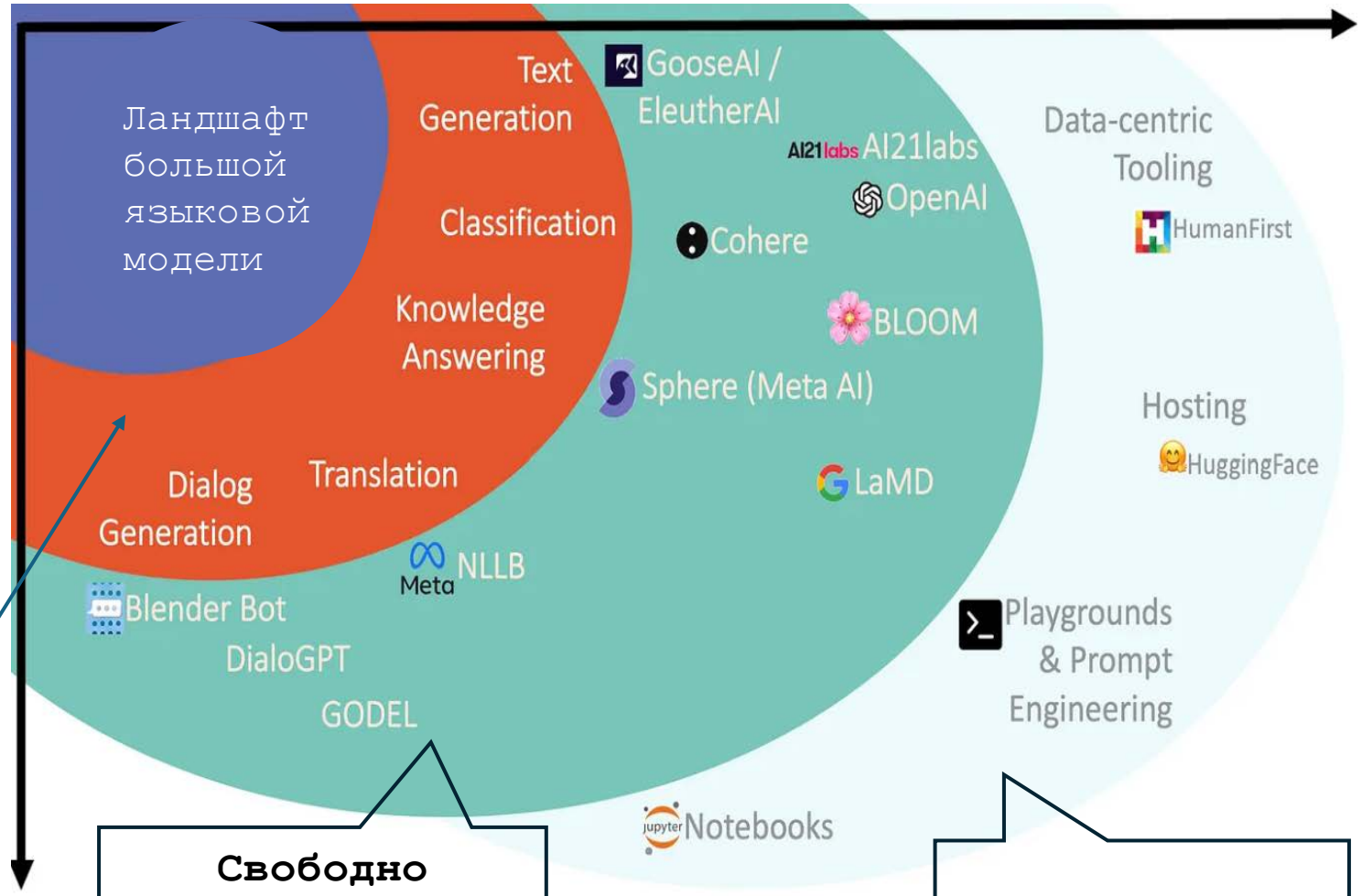
Группы высокого уровня по модернизации официальной статистики (HLG-MOS)



Ландшафт Больших языковых моделей: понятие, примеры

Большие языковые модели (LLM) – это класс искусственного интеллекта, который может понимать, интерпретировать и генерировать тексты

- Классификация
- Генерация ответа
- Генерация текста
- Перевод
- Генерация знаний (NLP)



Свободно распространяемые

Коммерческие

«Нет сомнения, что LLM будут играть важную роль в работе статистических организаций в будущем»

High-Level Group for the Modernisation of Official Statistics

modernstats by HLG - MOS

ДВЕ ОСНОВНЫЕ ГРУППЫ ЗАДАЧ

Распространенная практика в официальной статистике

<u>ПЕРВАЯ ГРУППА ЗАДАЧ</u>

Регулярные рабочие задачи: написание электронных писем и протоколов совещаний

<u>ВТОРАЯ ГРУППА ЗАДАЧ</u>

Перевод с SAS на R, обновления статистической системы классификации, создание отчетов, поиск данных на основе естественных языков и редактирование метаданных

РИСКИ: этические проблемы, правовые последствия (например, авторское право) и общая неосведомленность работников официальной статистики и низкая статистическая грамотность пользователей

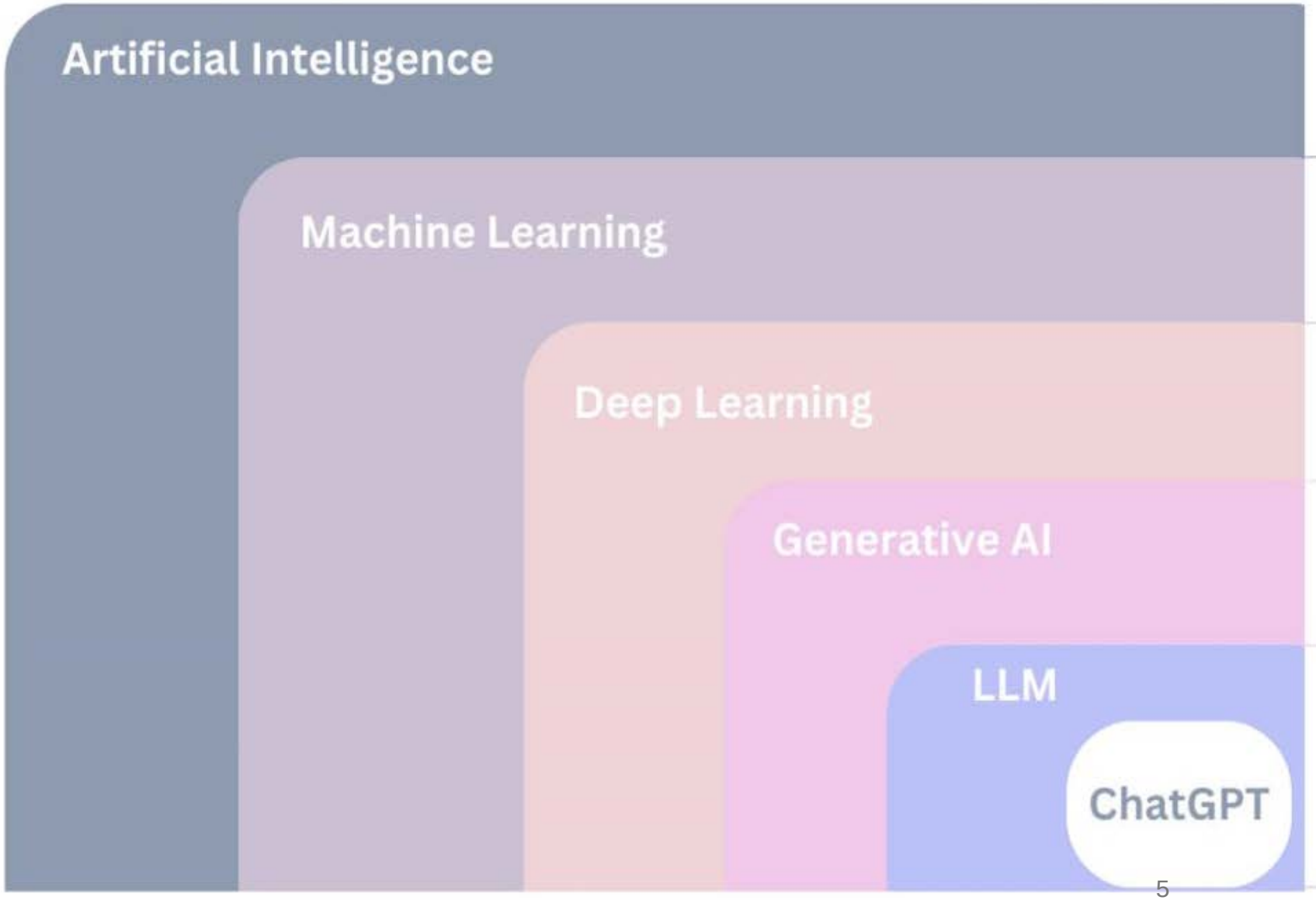
LLM в первую очередь предназначены для задач обработки естественного языка.

Основная функция: создание и понимание текста, похожего на человеческий.

Генеративный ИИ (GenAI или GAI) — это искусственный интеллект, способный генерировать текст, изображения, видео или другие данные с использованием генеративных моделей.

Модели генеративного ИИ изучают ***закономерности и структуру входных*** обучающих данных, а затем ***генерируют новые данные***

GENERATIVE AI AND LLM IN AI SPACE



1. Коммуникации: составление электронных писем, планов и отчетов, предоставление предложений по их содержанию, форматированию и генерации самого текста

2. Мозговые штурмы и генерация идей

3. Управление проектами и планирование. Автоматизация планирования задач, оптимизация распределения ресурсов на основе исторических данных и требований проекта. LLM облегчают управление встречами, автоматизируя создание повесток дня встреч и предлагая темы для обсуждения в соответствии с предопределенными целями или последними обновлениями.

(А) РЕКОМЕНДАЦИИ HLG-MOS (UNECE) ПО ИСПОЛЬЗОВАНИЮ LLM В ОФИЦИАЛЬНОЙ СТАТИСТИКЕ:

для целей управления и коммуникаций

7.Генерация изображения. Вместо того, чтобы покупать стоковые изображения, статистические организации могли бы использовать LLM для создания изображений, которые будут использоваться в статистических работах

5. Презентации- генерация содержимого слайдов, разработка тезисов с эффективным «тоном» презентации: настройкой для разных аудиторий

4. Перевод с/на другие языки документов, чувствительных к контексту

1. Дизайн опросов (GSBPM, подпроцесс 2.3) : проектирование опросов и анкет, разработка вопросов, форматы и формулировки, которые с большей вероятностью дадут точные ответы

2. Классификация и кодирование (подпроцесс GSBPM 5.2): автоматическая сортировка текстовые данные по предопределенным категориям или меткам

3. Проверка и редактирование данных (подпроцессы GSBPM 5.3 и 5.4): оптимизация задач очистки и предварительной обработки путем выявления и исправления ошибок данных, пропущенных значений и несоответствий

(Б) РЕКОМЕНДАЦИИ HLG-MOS (UNECE) ПО ИСПОЛЬЗОВАНИЮ LLM В ОФИЦИАЛЬНОЙ СТАТИСТИКЕ:
для целей повышения эффективности статистического производства и качества предоставления услуг

7. Помощь в **кодировании и переводе между языками программирования**

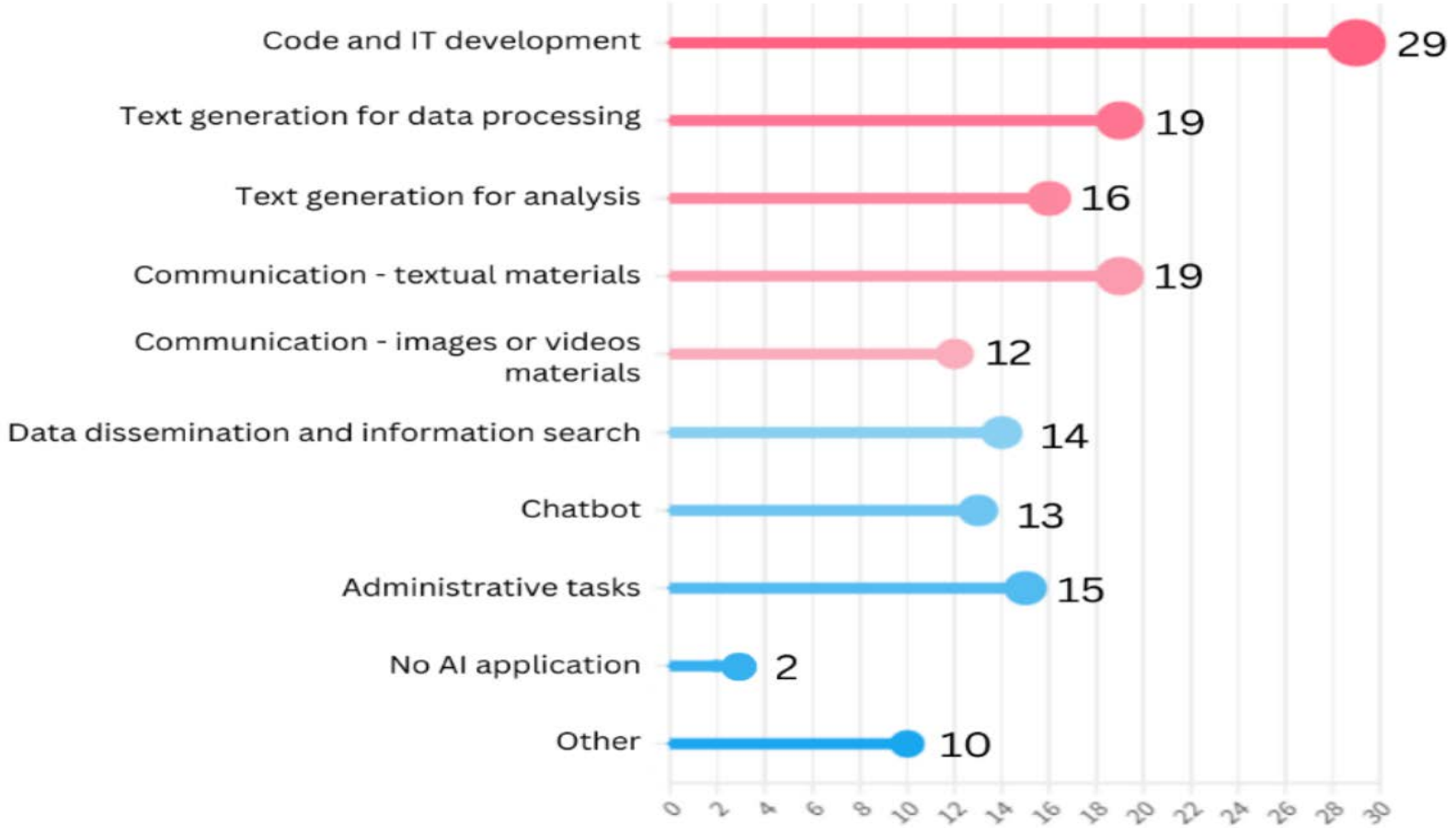
5.Производство продуктов для распространения (подпроцесс GSBPM 7.2): LLM могут генерировать текстовые описания таблицы или рядов чисел

6. Редактирование метаданных с помощью LLM

ОТВЕТ НАЦИОНАЛЬНЫХ СТАТИСТИЧЕСКИХ ОФИСОВ ЕВРОПЫ НА ВОПРОС ***HIGH-LEVEL GROUP FOR THE MODERNISATION OF OFFICIAL STATISTICS (HLG-MOS):***

КАКИЕ ОБЛАСТИ ПРИМЕНЕНИЯ ИСПОЛЬЗУЮТ ГЕНЕРАТИВНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ВАШЕЙ ОРГАНИЗАЦИИ (МНОЖЕСТВЕННЫЙ ВЫБОР)?

«Использование» включает как экспериментальные, так и производственные функции. 2024 г.



Зарубежный опыт применения LLM и больших генеративных систем в официальной статистике

Statistics Canada

Формирование отчетов с использованием LLM (Статистическое управление Канады)

Bank for International Settlements BIS

Редактирование метаданных с использованием GPT

	Highly impactful	Moderately impactful	Slightly impactful	Not impactful at all	Not sure	Average score
Data collection and processing	6	17	15	1	2	2,72
Data analysis	8	17	13	3	0	2,73
Dissemination and communication	13	16	9	2	0	3,0
Coding and IT development	21	15	4	0	1	3,43
Other administrative tasks	8	14	12	3	4	2,73

Оценка национальными статистическими органами Европы влияния генеративного ИИ на работу статистических организаций в ближайшие 2-3 года. Опрос *HLG-MOS, 2024 г.*

РОССТАТ ПРЕДСТАВИЛ СТРАТЕГИЮ РАЗВИТИЯ ГОСУДАРСТВЕННОЙ СТАТИСТИКИ ДО 2030 ГОДА

17 сентября состоялась стратегическая сессия, посвященная развитию отечественной статистики до 2030 года https://rosstat.gov.ru/folder/313/document/244701

Председатель Правительства РФ Михаил Мишустин:

- *«Важно, чтобы люди могли пользоваться достоверной и проверенной информацией, несмотря на непрерывный рост количества источников и типов данных.*

Для этого особое внимание мы уделяем системам анализа информации на базе искусственного интеллекта.

С их помощью можно получать более точные сведения и анализировать их в режиме реального времени. Это даёт возможность быстро принимать решения»

- *https://rosstat.gov.ru/folder/313/document/244701*

Для подтверждения целесообразности использования LLM в работе с метаданными официальной статистики проведен эксперимент по кластеризации методических указаний по статистике инвестиций, представленных на сайте Статкомитета СНГ, с использованием методов Text mining

Text mining –

- **процесс преобразования неструктурированного текста в структурированный формат для выявления значимых закономерностей и логических связей, которые невозможно получить непосредственно из чтения текста**

Text mining – входит в комплекс инструментов LLM и GenAI

Цель эксперимента: установить согласованность методик, представленных на сайте Статкомитета СНГ, на основе кластеризации их текстов на базе латентно-семантического анализа

Исходные материалы – методики по формированию показателей статистики инвестиций инвестиций в основной капитал, представленные на сайте Статкомитета СНГ: «ОБЩИЙ РАЗДЕЛ»

Методологические материалы НСС стран СНГ

Азербайджан

- Short methodological explanations

Беларусь

- Методика по расчету общего объема инвестиций в основной капитал и индекса физического объема инвестиций в основной капитал
- Методика по расчету статистического показателя Инвестиции в основной капитал за счет иностранных источников

Казахстан

- Методика по формированию показателей статистики инвестиционной деятельности
- Методика по определению объемов инвестиций в основной капитал с учетом скрытой и неформальной деятельности

Кыргызстан

- Методологическое положение по статистике инвестиций

Молдова

- Investments in fixed assets

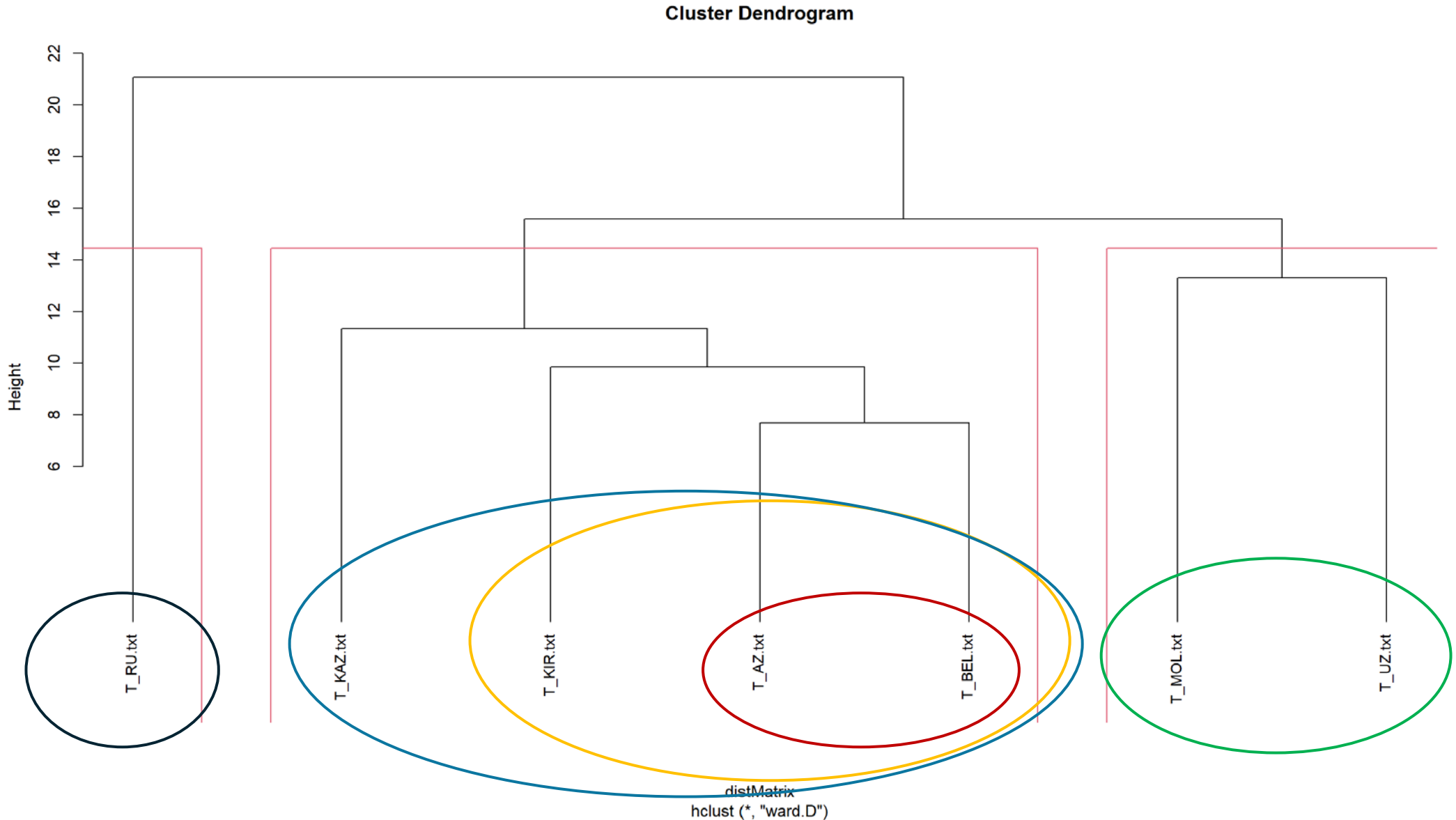
Россия

- Официальная статистическая методология определения инвестиций в основной капитал на федеральном уровне
- Официальная статистическая методология определения инвестиций в основной капитал региональном уровне
- Указания о порядке расчета индексов-дефляторов и индексов физического объема инвестиций в основной капитал

Узбекистан

- Методические положения по формированию общего объема инвестиций в нефинансовые активы

Результаты кластеризации методик по статистике инвестиций методами text mining



Выводы и направления дальнейших исследований

1. Применение больших языковых моделей (LLM) и генеративных систем ИИ (GenAI) – перспективное направление развития официальной статистики, обеспечивающее не только оптимизацию внутренних организационных процессов национальных статистических служб, но и повышение эффективности и качества производства и распространения официальной статистической информации
2. Необходима разработка методических рекомендаций для стран СНГ по применению больших языковых моделей (LLM) и генеративных систем ИИ (GenAI) в официальной статистике на основе имеющегося опыта стран СНГ, опыта других стран и международных сообществ
3. Эксперимент с применением технологий Text mining подтвердил целесообразность развития интеллектуального анализа текстовых данных и LLM для обеспечения согласованности метаданных официальной статистики стран СНГ
4. Развитие данных направлений должно сопровождаться обеспечение решения этических вопросов и вопросов сохранения национальной специфики производства статистической информации

Спасибо за внимание!

Zarova.EV@rea.ru