**Дорогие участники и участницы!**

Спасибо за то, что выбрали этот проект. Мы надеемся, что Вы получите удовольствие от его выполнения вместе со всеми.

В этом году каждый проект – сложная и комплексная задача. Не все шаги проекта из представленных могут быть решены в срок. И это нормально! Сам путь, который Вы проделываете может быть даже важнее, чем сам результат. Как часто говорят на “[ПАНДАНе](https://pandan.eusp.org/)”, отсылаясь к старой японской мудрости: “У студента нет цели, есть только путь.”

*А отсутствие подтверждения поставленных гипотез – тоже крутой результат!*  
  
Важно помнить, что после мастерской ничего не кончается. Вы можете также продолжать работу над проектом оставшимися энтузиастами. И мы – кураторы мастерской, будем рады в этом посильно помочь.

***Удаленные помощники***

Помимо кураторов Вам могут помочь удаленные помощники. Они – наши главные партнеры, предоставившие данные и идею проекта. Вы можете с ними пообщаться дистанционно. Они могут ответить на вопросы, касающиеся содержания данных, а возможно дать совет по поводу решения какого-то этапа (если Вы вдруг заблудились).

Для данного проекта удаленным помощником будет [Руслан Иванов](https://eusp.org/people/ruslan-vitalevich-ivanov), бэкенд-разработчик, Научной Лаборатории «Искусство и искусственный интеллект». Руслан также обучался на ДПО “ПАНДАН” в Европейском Университете. С ним можно связаться через Telegram: [@Rusl\_Ivanov](https://t.me/Rusl_Ivanov)

Хотим подметить, что удаленные помощники – не роботы и не БЯМы. У них есть свой график и распорядок основной деятельности. Они не всегда смогут ответить на возникшие вопросы в течении нескольких минут. Учитывайте это! Мы просим относиться ко времени и пространству других людей с уважением 🙂

***Принципы взаимопомощи***

Наша мастерская – отнюдь не про конкуренцию. Мы соревнуемся сами с собой в том, чтобы сделать себя лучше, а знания ещё острее. Хотим напомнить, что спрашивать – это нормально. Делиться – это нормально. Не стесняйтесь спрашивать преподавателей. Не смущайтесь просить помощи у других команд. Эти проекты – в том числе про обучение взаимодействию.

***Важные сроки***

Если Вы понимаете, что этот проект – “совсем не моё”, Вы можете поменять его до ***9 июля включительно***. Просим не делать это сюрпризом для остальной команды, а обсудить свое решение внутри коллектива и **обязательно** известить об этом координатора (обязательно получив его разрешение на перемещение).

Нам важно сохранить баланс каждой команды. Все же, у каждого проекта есть [свои лимиты](https://docs.google.com/spreadsheets/d/1SaDWNloBtkOGnK1bTZFgj-5ikbflW14xd3_N86n9pGw/edit?usp=sharing) (см. информацию в “Справке” по ссылке).

**Глоссарий:**

* **id\_hash** - уникальный идентификатор человека
* **gender** - пол
* **age** - возраст
* **location** - место пропажи (нас. пункт)
* **status** - статус поиска
* **date\_search** - дата начала поисков
* **date\_of\_loss** - дата пропажи человека
* **last\_search\_date** - последняя дата поиска
* **search\_period** - срок поиска человека

***0. Задания ниже – это не обязательные пункты, а желательные ориентиры. Вы всегда можете придумать что-то свое, если найдете какие-то инсайты в данных или захотите исследовать авторскую гипотезу.***

**1. Очистка данных. В этом наборе есть сильные зашумления, от которых нужно избавиться. Проведите диагностику данных.**

В следующих переменных есть явные проблемы: *gender; date\_search, date\_of\_loss, last\_search\_date; location; age; status.*

1. Визуализация проблем в данных:
   * Какие категории являются наименее наполненными? Есть ли способы решить эту проблему?
   * Есть ли в данных аномалии?
2. Обработка поля gender (пол):
   * Привести все значения к единому формату: М, Ж или Unknown (если пол не указан).
   * Нормализовать возможные опечатки (например, "Муж.", "мужской", "жен").
   * ***Попробуйте автоматически определить пол человека по имени, используя библиотеки (например, gender-guesser или pymorphy2 для русских имен).\****
3. Исправление дат (date\_search, date\_of\_loss, last\_search\_date):
   * Проверить корректность дат (например, date\_search не может быть позже last\_search\_date).
   * Если дата пропажи (date\_of\_loss) отсутствует, но есть дата начала поиска (date\_search), можно предположить, что пропажа произошла незадолго до этого.
   * Исправить явные аномалии (например, даты из будущего, несуществующие даты).
   * Убедиться, что search\_period корректно вычисляется как разница между last\_search\_date и date\_search.
4. Заполнение пропущенных значений:
   * Если location не указан, попробовать извлечь его из текстовых описаний (если такие есть в исходных данных).
   * Для age можно попробовать восстановить данные, если есть упоминание возраста в тексте.
5. Проверка статуса поиска (status):
   * Нормализовать возможные варианты ответа в переменной.

**2. Требуется дособрать данные о пропавших людях с форума. Соберите максимальное количество данных** [**с форума**](https://lizaalert.org/forum/) **за недостающие периоды.**

**3. Проведите анализ закономерностей в данных.**

#### 1. Рассчитать длительность поиска в разных регионах.

#### 2. Влияние демографических факторов на срок поиска

* Сравнить среднюю продолжительность поиска между мужчинами и женщинами.
* Проанализировать зависимость срока поиска от возраста, выделив ключевые возрастные группы.
* Выявить регионы с наиболее выраженными различиями.
* Ваш вариант

#### 3. Визуализируйте свои результаты

#### 4. Анализ сезонных закономерностей

* Исследовать распределение случаев пропажи по месяцам и сезонам.
* Определить, существуют ли регионы с выраженной сезонной динамикой.

*\*5. Картография*

* *Постройте карту, визуализирующую пропажи по регионам России*

**4. Постройте классификатор, определяющий вероятность нахождения человека (в любом статусе) в зависимости от его демографических характеристик, географического района, местности и пр. Постройте альтернативные классификаторы для разных итогов нахождения (Мертв, Жив). Какой из них работает лучше?**

* Является ли модель достаточно точной? Можно ли её отладить с помощью анализа отдельных сущностей из описания пропавшего человека?

**5. Сгенерируйте данные-описания пропавших людей (любым способом). Повторите паттерн написания с форума. Проанализируйте, работает ли это на сгенерированных данных? Насколько хорошо?**

* Если алгоритм Ваших действий работает на оригинальных данных, включите синтезированные данные в ваш пайплайн.

На этом этапе Вам могут помочь эти материалы (но можно ориентироваться на прошедшие занятия):

[MjdMahasneh/LLM-Fine-Tuning-With-Hugging-Face-Transformers: In this tutorial, we’ll walk you through the steps to fine-tune an LLM using the Hugging Face transformers library, which provides easy-to-use tools for working with models like GPT, BERT, and others. we’ll also provide a code demo for fine-tuning GPT-2 (a smaller version of GPT-3) on a custom text dataset.](https://github.com/MjdMahasneh/LLM-Fine-Tuning-With-Hugging-Face-Transformers)