**К этапу №1 "Сбор данных"**

На первом этапе вам предлагается при помощи API YouTube автоматическим образом собрать базу данных о релевантных видео одним из трёх путей:

1. видео с конкретного(ых) канала(ов)
2. видео по конкретному поисковому запросу
3. видео из конкретного региона за определённый промежуток времени

Предложенные пути можно комбинировать и дополнять. Выбранный путь должен соответствовать поставленному исследовательскому вопросу. В итоговой базе должны соблюдаться следующие требования:

1. количество собранных видео – не менее 1000, при этом в базе не должно быть повторяющихся видео.

2. для каждого видео должны быть представлены следующие характеристики: название, описание, дата публикации, длительность, название канала, количество просмотров, лайков, дизлайков, комментариев.

3. в базе не должно быть лишних (т.е. дублирующих, технических и тех, которые впоследствии не предполагается анализировать) характеристик

4. база должна быть оформлена в таблицу Excel.

**Требования и рекомендации к этапу №2 "Первичный анализ данных и text mining"**

Задача на парную связь. В этой задаче могут участвовать переменные любых типов шкал. Хотя бы одна из этих переменных должна быть выведена из исходных переменных выдачи API YouTube, а не взята «как есть». Ожидаем следующий порядок действий: сформулировать содержательную задачу; выбрать и создать релевантные переменные, описать каждую из них с т.з. типа шкалы, репрезентрирующего значения и однородности; по возможности, сформулировать содержательную гипотезу; обосновать метод; при необходимости подготовить переменные (удалить выбросы, сгруппировать категории); применить метод и проинтерпретировать результаты; по возможности, применить визуализацию.

Задача на text mining. В этой задаче участвуют тексты (1-5 тысяч). Содержательно задача может относиться к одному из [четырёх типов](https://docs.google.com/document/d/1sFRdw2FR-GIQ3uhB5Ykl3JXmfkfqsAM8VqzCy94frw8/edit?usp=sharing): 1. Выявить среди собранных документов (текстов) наполненные и для них – высокочастотные токены. 2. Выявить среди собранных документов нерелевантные. 3. Выявить во множестве документов обозримое число интерпретируемых топиков (тем). 4. Выявить сентимент (эмоциональную нагрузку) собранных документов. Порядок решения задачи: сформулировать содержательные задачу, удалить «мусорные» (в контексте содержания задачи) символы, лемматизировать слова (привести слова к начальной лексической форме), удалить стоп-слова (в контексте содержания задачи), векторизовать документы. Представить результаты рекомендуем таким образом: 10 примеров искомых документов и/или топиков в связке с релевантными токенами; показатели качества построенной модели (если уместно); интерпретация.