# Задание 1 (2 балла)

## Disk Usage

### Дедлайн - 17.03.2022 20:59

Необходимо написать аналог утилиты du с опцией -sh

Предполагается, что утилита всегда работает с файлами и директориями, в поддиректориях нет ссылок на другие файлы, все файлы при этом могут быть прочитаны пользователем.

Можно пользоваться только командами stat (или ls -l) и awk

# Задание 2 (2 балла)

## Train val split

### Дедлайн - 17.03.2022 20:59

Напишите скрипт, который позволяет разбить датасет в формате csv на train и val выборку.

Можно протестировать работу на датасете:

<https://www.kaggle.com/c/titanic>

Скрипт принимает следующий набор параметров:

* --input … (путь к датасету)
* --train\_ratio … (доля объектов в обучающей выборке)
* --shuffle (Флаг есть, если необходимо перемешать датасет)
* --train\_faile … (путь к train данным)
* --val\_faile … (путь к val данным)

# Задание 3 (2 балла)

## Parallel Dataset Processing

### Дедлайн: 24.03.2022 20:59

Необходимо написать скрипт, который будет параллельно выкачивать данные для какого-то датасета.

В аргументы скрипта необходимо передать:

* количество worker-ов (--num\_workers)
* файл (--input\_file)
* столбец, в котором располагаются ссылки на данные (--links\_index)
* папка для сохранения данных (--output\_folder)

Все аргументы должны быть именованными

Ссылка на датасет - <https://drive.google.com/file/d/1EfRc2RLVdwWlXWz3nDIBEv_EvvOMd9ip/view?usp=sharing>

# Задание 4 (3 балла)

## Bioinformatics Dataset Processing

### Дедлайн 24.03.2022 20:59

В pipeline задач биоинформатики данные занимают десятки и сотни гигабайт, поэтому все данные хранятся только в архивном варианте (промежуточные варианты не хранятся в чистом формате).

Вам необходимо будет сделать утилиту, которая в зависимости от опций делает одну из следующих команд:

* сохраняет небольшой стартовый участок файла (1 балл)
* оставляет только те прочтения, которые имеют качество прочтения, не менее определенного порога (2 балла)

Качество прочтения - минимальное качество для каждого нуклеотида внутри прочтения. Нуклеотид - одна из 4 букв A, C, G, T

Формат файла FASTQ: <https://en.wikipedia.org/wiki/FASTQ_format>

Пример файла FASTQ можно скачать здесь:  
<https://digitalinsights.qiagen.com/downloads/example-data/>

Прежде чем выполнять задание на данных сэмплах - не забудьте их зашифровать в tar.gz формат.

# Задание 5 (6 баллов)

## JSON Dataset Processing

### Дедлайн: 31.03.2022 20:59

Скачайте датасет COCO по ссылке: [2014 Train/Val annotations [241MB]](http://images.cocodataset.org/annotations/annotations_trainval2014.zip) - используем instances датасеты

Описание датасета доступно по ссылке: <https://cocodataset.org/#download>

Необходимо сделать следующие действия при помощи bash и утилиты jq:

* научиться выделять все изображения по категории данного класса (1 балл)
* научиться отфильтровывать мелкие изображения: размер которых по каждой размерности менее определенного порога (1 балл)
* фильтровать датасет по количеству объектов изображений в определенной категории (2 балла)
* фильтровать изображения, в которых размер маски определенного класса более определенного порога по количеству пикселей (2 балла)