# Домашнее задание №5

| | |
|---|---|
| Дисциплина | Наука о данных для юристов |
| Тема | Тема 1. Введение в Python |
| Форма проверки | **Задание с индивидуальной проверкой** |
| Имя преподавателя | Кирилл Сиротинский |
| Время выполнения | 2 часа |
| Цель задания | Научиться работать с файлами, читать из них информацию и записывать информацию в файлы; научиться работать с регулярными выражениями |
| Инструменты для выполнения ДЗ | • Python<br>• GitHub |
| Правила приёма работы | Зарегистрируйтесь на https://github.com/.<br><br>Создайте публичный репозиторий для приёма домашних работ. Наименование репозитория выберите по следующему шаблону: HSE_Ivan_Ivanov. Прикрепите ссылку на репозиторий.<br><br>Для каждого домашнего занятия код загружается в отдельные папки: lesson 1 для ДЗ №1 и 2, lesson 2 для ДЗ №3 и 4 и т. д.<br><br>Итоговое задание вы загрузите в папку final |
| Критерии оценки | Оценка задания производится преподавателем на основе следующих критериев:<br>• по итогу работы скрипта в папке сохраняется файл *traders.csv* с информацией об ИНН, ОГРН и адресе по списку из файла *traders.txt* — 4 балла;<br>• по итогу работы скрипта в папке сохраняется файл *emails.json*, в котором собраны все адреса электронной почты с привязкой к ИНН публикатора — 4 балла;<br>• весь код оформлен в соответствии с общепринятыми правилами (PEP 8), поиск адресов электронной почты во второй части задания реализован не только в разделе с основным текстом сообщения — 2 балла.<br><br>Максимально можно получить 10 баллов |
| Дедлайн | **11.03.2024, 23:59** |

# Задание

1. Найдите информацию об организациях.
    a. Получите список ИНН из файла **traders.txt**.
    b. Найдите информацию об организациях с этими ИНН в файле **traders.json**.
    c. Сохраните информацию об ИНН, ОГРН и адресе организаций из файла **traders.txt** в файл **traders.csv**.
2. Напишите регулярное выражение для поиска email-адресов в тексте.

Для этого напишите функцию, которая принимает в качестве аргумента текст в виде строки и возвращает список найденных email-адресов или пустой список, если email-адреса не найдены.

Используйте датасет на 1 000 сообщений из Единого федерального реестра сведений о банкротстве (ЕФРСБ) для практики.
Есть датасеты и побольше:

- датасет на 10 000 сообщений,
- датасет на 100 000 сообщений.

Если компьютер слабый, ограничьтесь самым маленьким.

Текст сообщений можно найти по ключу **msg_text**.

Найдите все email-адреса в датасете и соберите их в словарь, где ключом будет выступать ИНН опубликовавшего сообщение publisher_inn, а в значении будет храниться множество set() с email-адресами. Пример:

```
{
“77010127248512”: {“name_surname@yandex.ru”, “name_surname@mail.ru”}
“77011235421242”: {“name_surname@yandex.ru”, “name_surname@gmail.com”}
…
}
```

Сохраните собранные данные в файл **emails.json**.