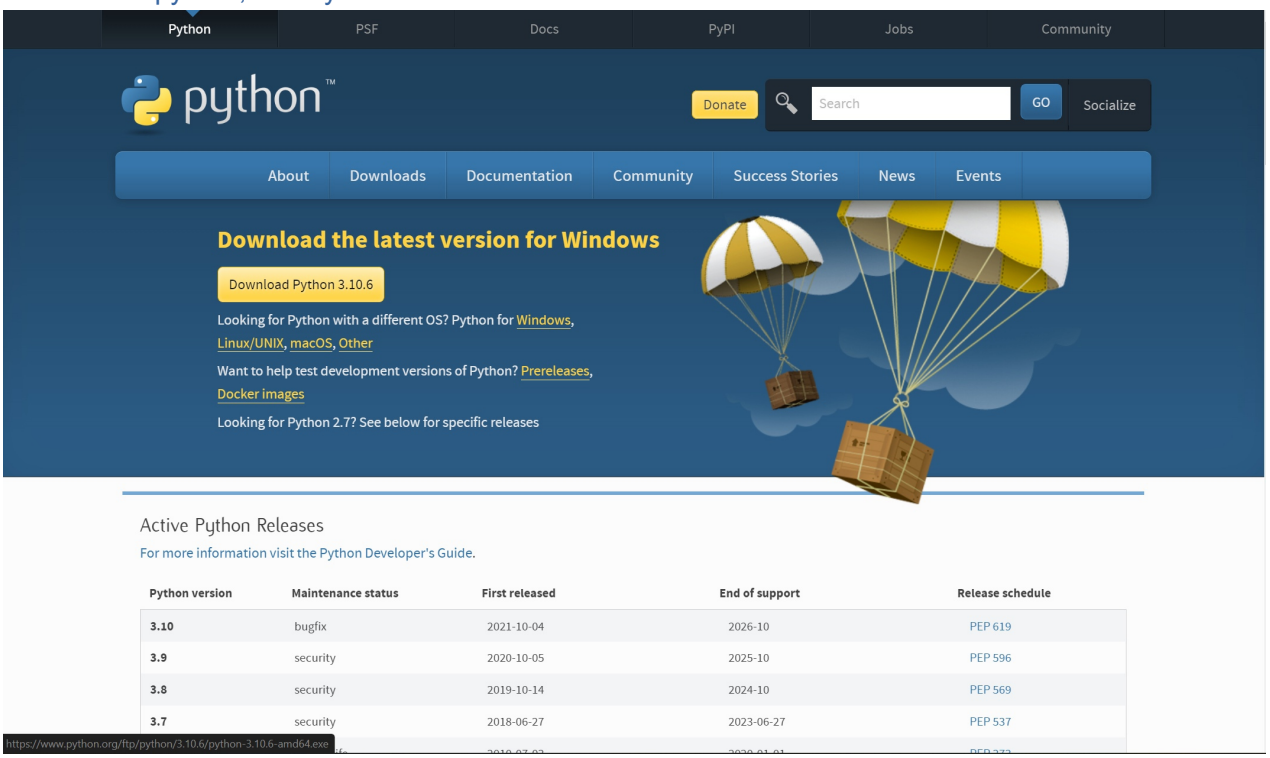
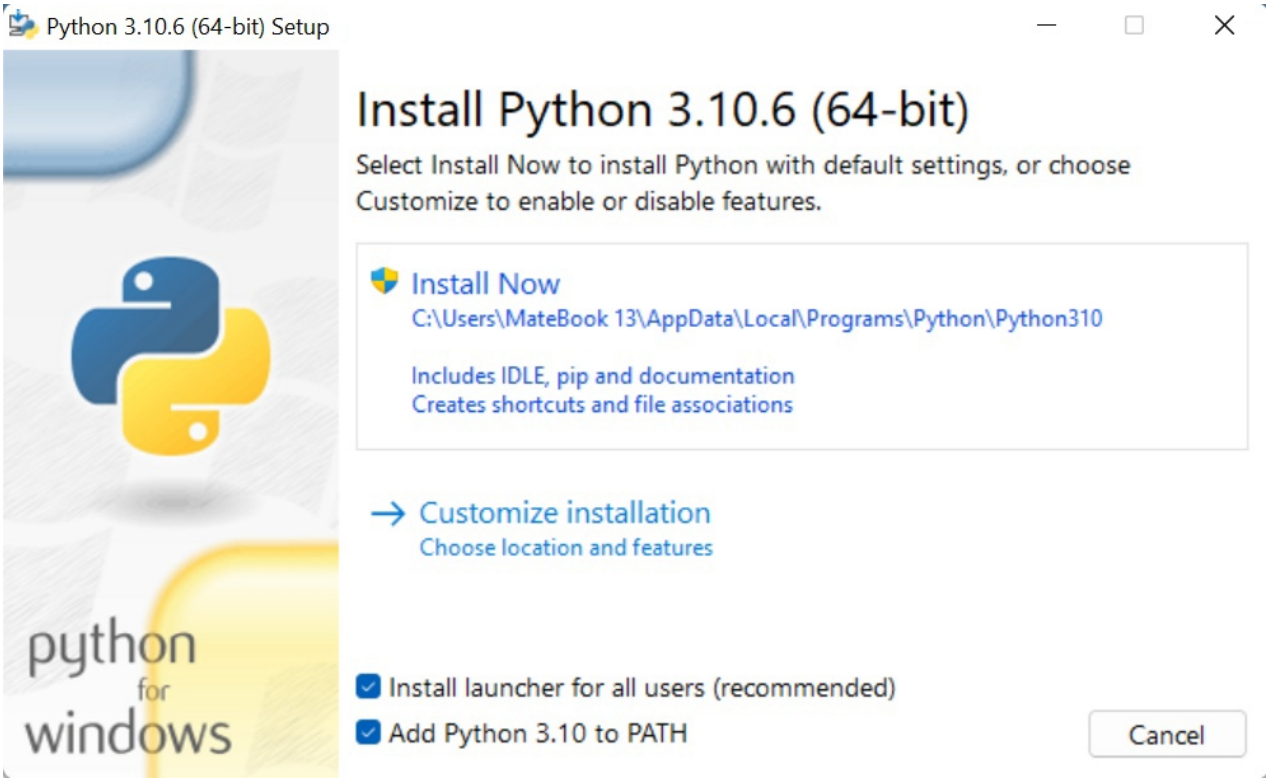
# Установка

## Шаг 1

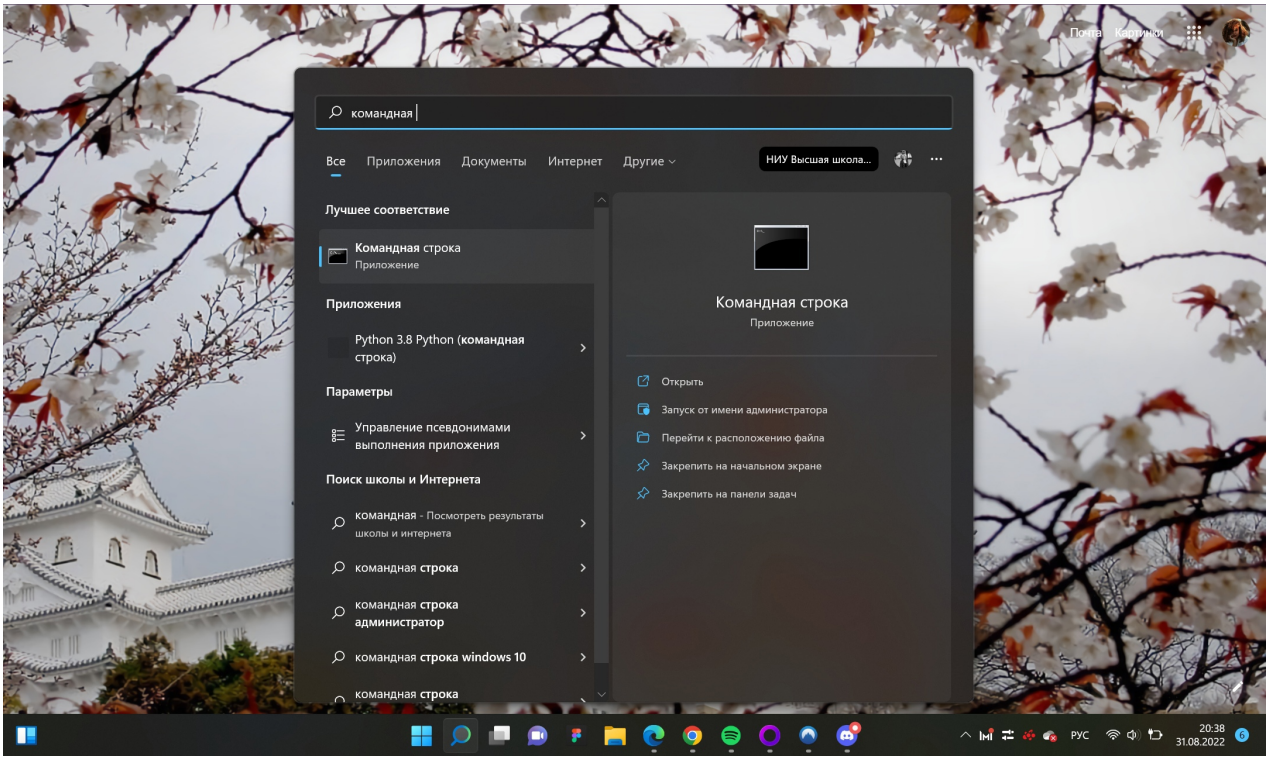
1. Скачиваем python, если у вас его нет.




2. Запускаем инсталятор и ставим все галочки как на скрине




3. Открываем командну строку (заходим в поиск и вводим 'командная строка', либо используем миллион других способов, которые легко загуглить)



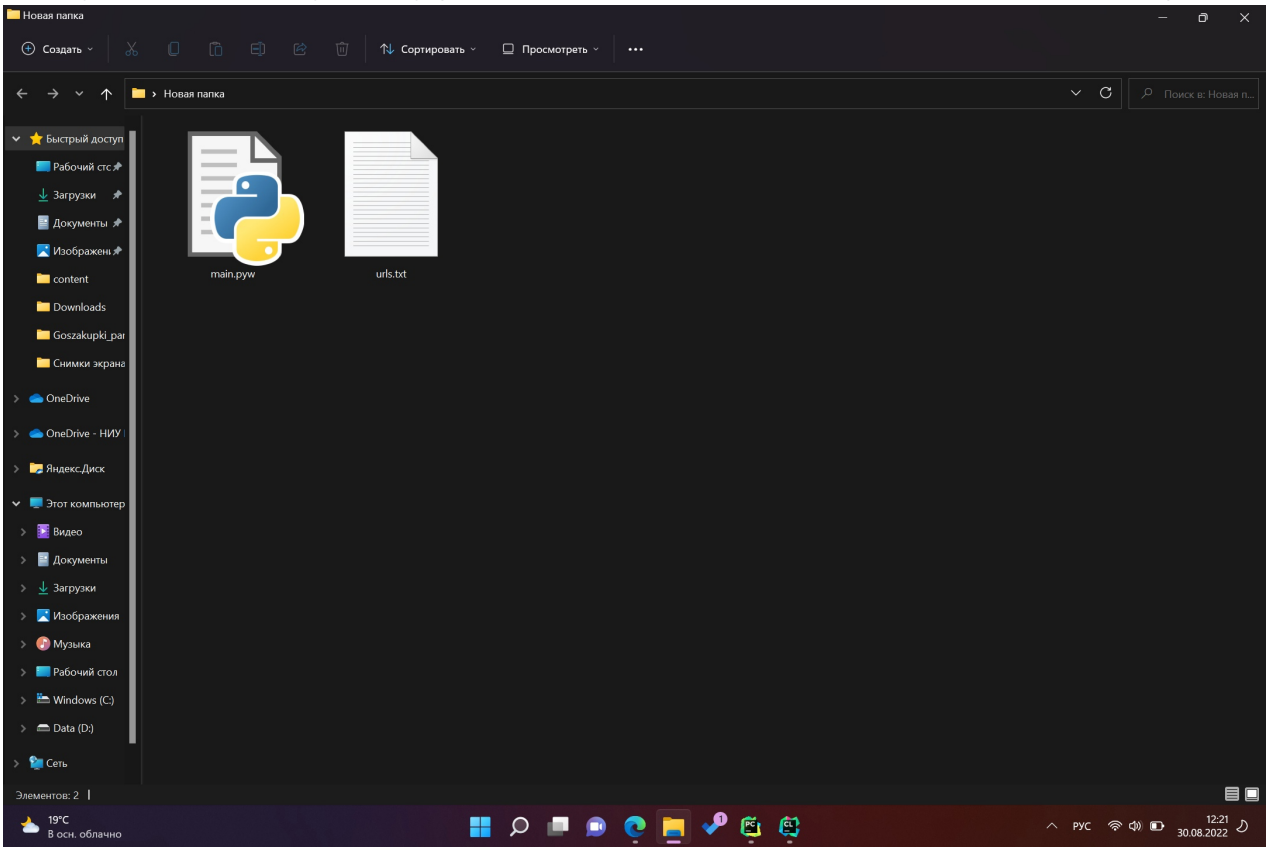

4. Вводим в командную каждую строку из тех, что можно увидеть ниже (ВАЖНО: Вводим построчно и после каждого ввода нажимаем Enter и ожидаем загрузки пакета. Затем повторяем действие еще 9 раз)

```
pip3 install requests
pip3 install python-docx
pip3 install aspose.words
pip3 install bs4
pip3 install docx
pip3 install py7zr
pip3 install pyunpack
pip3 install pandas
pip3 install openpyxl
pip3 install patool
```

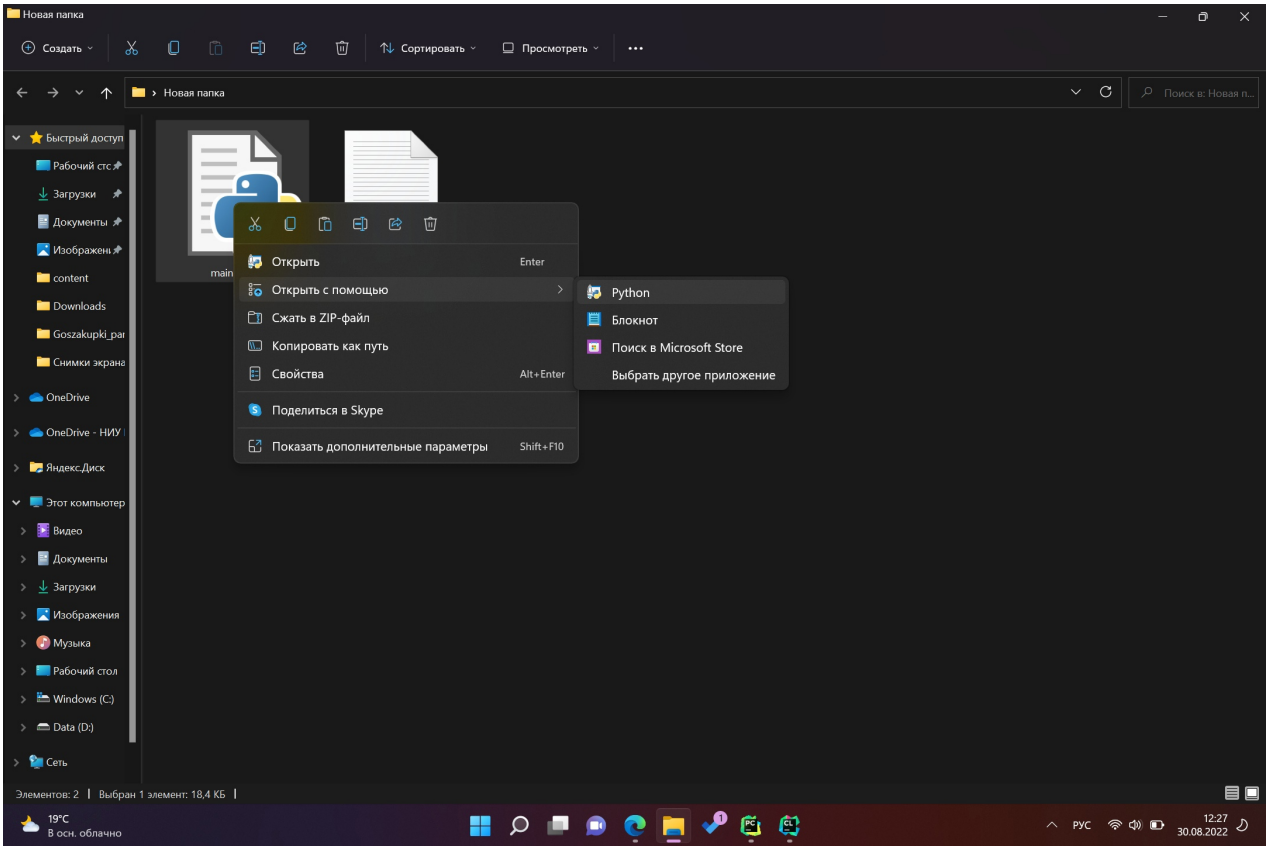
5. После того как у нас все установилось можно переходить к шагу 2

# Шаг 2

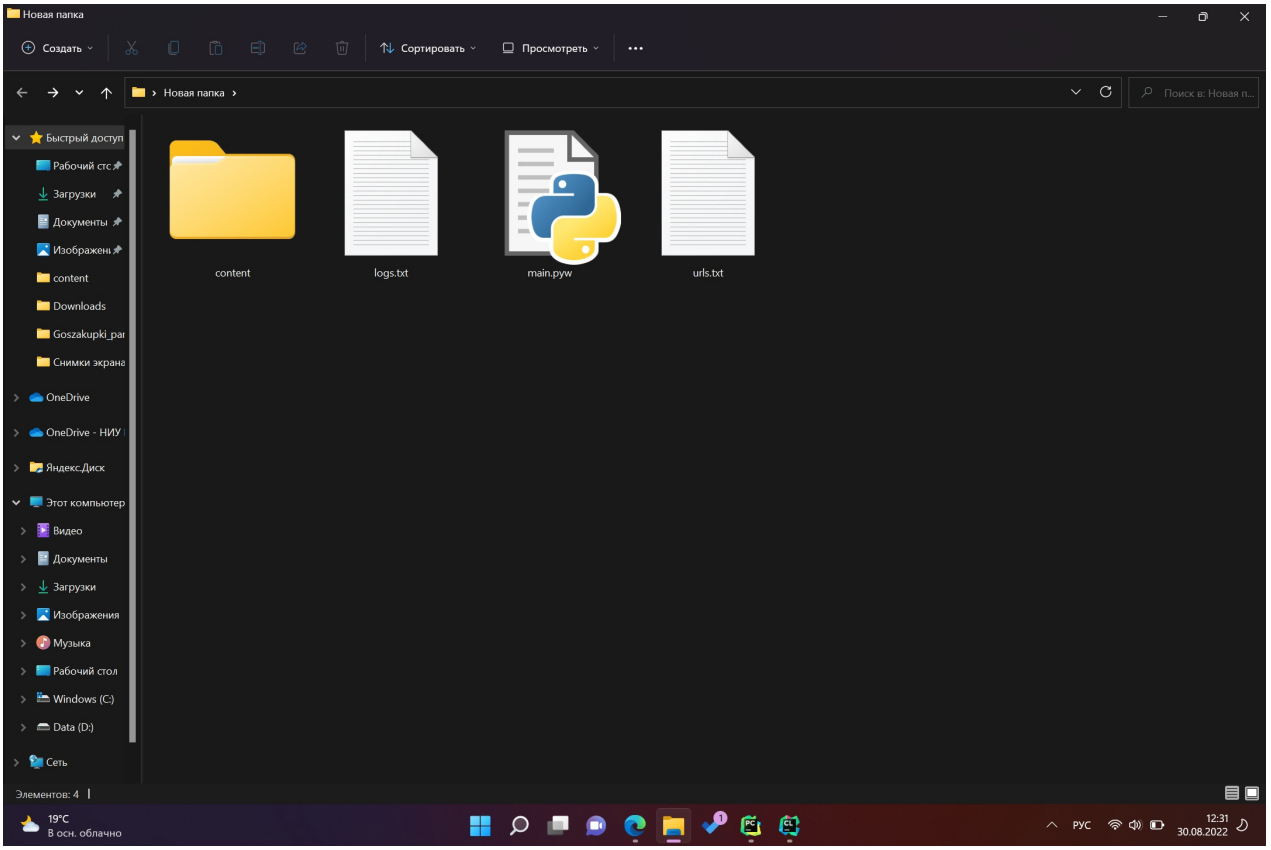
1. Копируем файлы в любую папку (наличие файла *urls.txt* обязательно, как и его название)



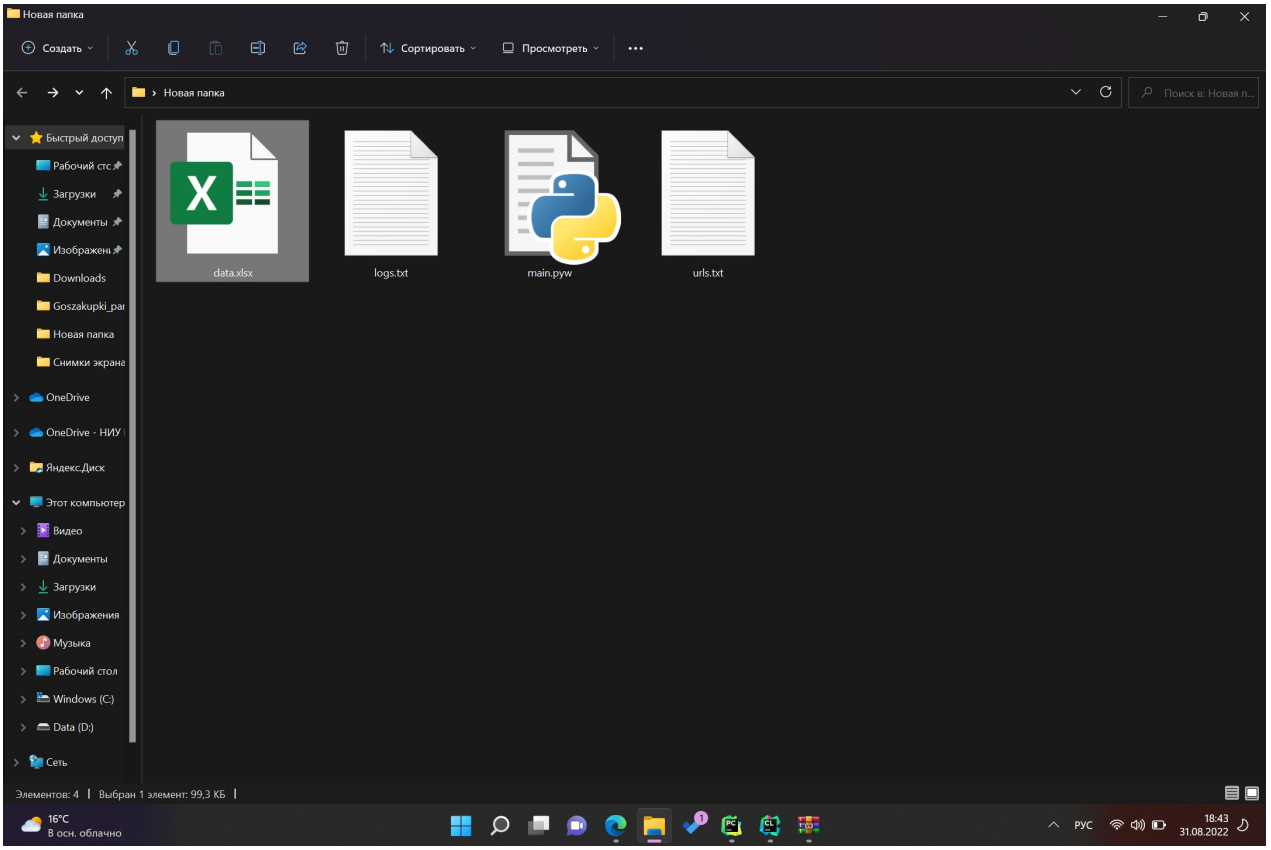

2. Открываем файл и ждем некоторое количество времени. Зависит от количества ссылок, которые нужно обработать, а также от степени нагруженности на госзакупки




3. После запуска появится файл *logs.txt*, в котором указаны некорректные ссылки(ссылки другого формата или отсутствие документов для загрузки), а также папка *content*, в которую скачиваются документы для дальнейшей обработки. Папку не нужно трогать и обижать потому что может что-то пойти не так.




4. После завершения обработки *content* удаляется вместе со всеми скачанными файлами и появляется файл *data.xlsx*, в котором находится итоговая информация




# To do

1. Не всегда корректно собирается информация с таблиц из-за наличия таблиц всех цветов и оттенков. Довольно трудно сделать что-то, что будет разом подходить под все.
2. Трудности с обработкой pdf формата
3. Дублирование информации (решаемо)
4. Нехватка ключевых фраз. Крайне редко, но замечал, что пропускается информация из-за другого расположения слов или наличие скобок с пояснением между фразами
5. Указать из какого документа берется данная нформация и выделить ключевой момент (Из-за того, что у госзакупок очень странно работает переадресация на скачивания документа, то имя документа может выглядеть вот так: doc04051320220608130025 или Æ«¼ 2 ÆÑσ¡¿тÑß¬áՈ táßГ∞.pdf из-за сбитой кодировки. Пока не ясно как это можно обойти, попробую что-то придумать)
6. Сделать какую-то визуализацию работы чтобы можно было следить за прогрессом обработки ссылок