**Скачивание отчётов**

В программе *download\_pages.py* получаем HTML-код страницы и оттуда получаем все значения *pk*= «…» так как все отчёты в формате .*pdf* открываются по адресу типа <http://unro.minjust.ru/Reports/93962201.pdf>, где 93962201 – значение атрибута *pk.*

После этого проходим по списку значений и пробуем скачать каждый отчёт.

Скрываем свой IP с помощью нескольких прокси, взятых с сайта [https://hidemy.name/](https://hidemy.name/ru/proxy-list/?ports=8080&type=h&anon=34&__cf_chl_jschl_tk__=a147353863fc5ec29617f7672bccea4eb12d0c81-1593062085-0-Abop3kQYexDQqlVcslQ2jCqCT1jhD6Zp7vVEdstjlPudsURG_tMlV726k2Z14sX-xrzO_te5i_Ss6FgEV4ptrEE3JhXKpy-zwVfnzGHeaCc6OmlPUYk3HhP9xSgbhSTkXSgQRiShLmKLBMNkulvFMLR2cTXaWten4j_osu4iqxAb0MXFm9SfmoaYINu3e-7jvqeHqOUFoWIFlMgLYLBCK6uzfflKm_98QURjlIMDK0D1FXQ0pTM9f6F_dtNZAG7s5hmPSAbnLLbmgf190os7gSvbpSDd_bO94bk-jJNJurQY2UiSjID5nzwsAGVXYmc7LoP2lxBnngj6JbYn38SSC6U#list). Для каждого IP создаём отдельную папку, куда будет сохраняться отчёт.

Номера отчётов, которые скачать не удаётся, сохраняем в файл *listfile.txt*.

**Парсинг отчётов**

Происходит в *pdf\_parse.py.* Каждый документ загружается, приводим его в вид строки при помощи модуля *fitz*. Далее в документе находим контрольные точки и запоминаем значения между ними в большой *DataFrame.*

**Импорт в БД**

Из большого общего *DataFrame* производим несколько других, которые будут являться таблицами в БД и записываем их в БД при помощи модуля *sqlalchemy.*

**Проектирование БД**

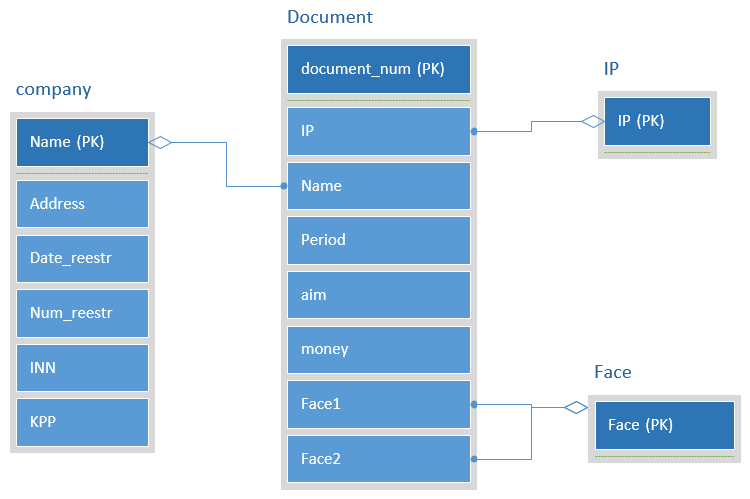


Рисунок 1 – ER-Диаграмма в нотации *IDEF1X*

Было выделено 4 сущности. Из-за разнообразности отчётов (отсутствия ИНН, различных записей названия одной и той же компании, различного написания даты и т. п.) было решено в сущности *Face* не выделять должность в отдельный атрибут и в *company* не использовать ИНН в качестве первичного ключа (много отчётов с незаполненным полем ИНН).

**Минусы**

- не удалось покрыть код программы unit тестами.

- при прерывании работы, программу нужно запускать с начала.

- присутствует некоторое дублирование в БД из-за разнообразности оформления отчётов.

- нет поддержки аутентификации по логину и паролю.

- в БД не удалось произвести необходимые связи между таблицами.