Бухгалтерская отчетность 1,6 млн российских компаний

Как получить и как использовать?

Евгений Погребняк

Version 0.0.1, January 29, 2019

Мы сделали интерфейс для удобного доступа к открытым данным корпоративной отчетности российских компаний (РСБУ) за 2012-2017 год. Пользователи могут получить эти данные в виде преобразованных CSV файлов, с которыми далее можно работать средствами pandas или R.

В преобразованных CSV файлах записаны заголовки с названиями столбцов (переменных). Размер исходных файлов, составляющий 0,7-2,3 Гб за один год, уменьшается примерно в два раза за счет удаления пустых и неиспользуемых столбцов.

Код для доступа к данным и примеры их использования находятся в репозитарии https://github.com/ru-corporate/sandbox. Исходные файлы находятся на сайте Росстата.

Как получить отчетность?

Установка пакета:

```
git clone https://github.com/ru-corporate/sandbox.git ①
cd sandbox
pip install -r requirements.txt
pip install . ②
```

① вместо git также можно скачать и распаковать zip файл

② устанавливаем `boo` локально

Загрузка данных за 2012 год:

```
import boo
boo.download(2012) ①
boo.build(2012) ②
df = boo.read_dataframe(2012) ③
```

① скачиваем исходные данные

② преобразуем их и записываем в новый файл (см. ниже)

③ читаем данные из преобразованного файла в фрейм `pandas`

Для преобразования отчетности используется следующий алгоритм:

1. ввести названия столбцов, отражающие смысл переменных бухгалтерской отчетности (например, счет "Основные средства", код `1150`, обозначение в файле `of`)
2. уменьшить размер файла за счет удаления пустых и неиспользуемых столбцов
3. привести все строки к одинаковым единицам измерения (тыс. руб.)
4. создать новые столбцы:
    - короткое название компании
    - код ОКВЭД разбить на три уровня

    - определить регион по ИНН
5. убрать дублирование данных по коду ИНН (~30-40 организаций)
6. указать тип переменных по столбцам, чтобы ускорить чтение данных

Дополнительная информация о загрузке и преобразовании данных находится в репозитарии ru-corporate/sandbox.

Как использовать?

Сферы применения отчетности:

- рейтинги крупнейших компаний и отраслевые рейтинги
- финансовая устойчивость и эффективность компаний
- оценка стоимости компаний, кредитный риск, банкротства
- оценка экономической концентрация в отрасли
- анализ налоговых поступлений и налоговой нагрузки
- инвестиционный процесс (кто инвестирует в расширение мощностей)
- региональная экономика и планирование (городской транспорт, ЖКХ, ТБО)

Что хотим показать?

- краткий код доступа к данным
- полноту данных
- разметку по отраслям
- крупнейшие холдинги
- красивые графики

Возможные разделы

- Карта экономики
- Использование в обучении финансам

Учим asciidoc

Расширения:

- Можем ли вставить таблицы из pd.DataFrame?
- asciidoctor-new-noteworthy-beyond

Базовые знания:

- asciidoc-syntax-quick-reference.adoc
- asciidoc_example.adoc