ЖУРНАЛ ИНЖЕНЕРНЫХ РАБОТ ПО ПРОЕКТУ "УСОВЕРШЕНСТВОВАННАЯ МОДЕЛЬ РИЧАРДСОНА"

План работ
1. Поиск, фильтрация, верификация, выбор данных.
2. Подготовка данных, конвертация, заливка в БД.
3. Нормализация структуры БД.
4. Предобработка данных (Data Preprocessing).
  - Очистка данных
  - Оптимизация данных:
    - Факторный анализ (Factorial analysis)
      - Метод главных компонент (Principal component analysis)
        (использовать библиотеки языка программирования Python).
5. Визуализация данных
  - в виде интерактивных таблиц
  - в виде диаграмм и графиков

1. Поиск, фильтрация, верификация, выбор данных.

CSV-файлы скачанные из источников:
  - ВВП с сайта Всемирного банка: vvp.csv
  - ИРЧП с сайта ПроООН: irchp.csv и др.
Полный список и адреса источников см. в БД в таблице source.

2. Подготовка данных, конвертация, заливка в БД.

Приведение таблиц скачанных с сайтов к виду пригодному к заливке в БД.

В Postgresql создана БД richard

Сырые данные (неочищенные и не нормализованные)

Создание буферной таблица для приема данных:

```
create table tmp (a0 varchar(255),
  a1 varchar(255), a2 varchar(255), a3 varchar(255),
  a4 varchar(255), a5 varchar(255), a6 varchar(255),
  a7 varchar(255), a8 varchar(255), a9 varchar(255),
  a10 varchar(255), a11 varchar(255), a12 varchar(255),
  a13 varchar(255), a14 varchar(255), a15 varchar(255),
  a16 varchar(255), a17 varchar(255), a18 varchar(255),
  a19 varchar(255), a20 varchar(255), a21 varchar(255),
  a22 varchar(255), a23 varchar(255), a24 varchar(255),
  a25 varchar(255), a26 varchar(255), a27 varchar(255),
  a28 varchar(255), a29 varchar(255), a30 varchar(255),
  a31 varchar(255), a32 varchar(255), a33 varchar(255),
  a34 varchar(255), a35 varchar(255), a36 varchar(255),
  a37 varchar(255), a38 varchar(255), a39 varchar(255),
  a40 varchar(255), a41 varchar(255));
```

Заливка данных в буферную таблицу
```
\copy tmp FROM '/home/an2k/py/richard/data/vvp.csv' DELIMITER ':' CSV
```


3. Нормализация структуры БД.

Созданы таблицы для 3-х измерений и собственно данных:

```
drop table source;
create table source (sid serial primary key, sname varchar(255), slink varchar(255));
create table year (yid serial primary key, yname varchar(255));
drop table country;
create table country (cid serial primary key, cname varchar(255));

drop table value;
create table value (
  vid serial primary key,
  vsid int,   -- id источника
  vyid int,   -- id года
  vcid int,   -- id страны
  val float   -- значение
);


\d source
                              Таблица "public.source"
 Столбец |          Тип           |                     Модификаторы
---------+------------------------+-----------------------------------------------------
```

```
 sid     | integer                | NOT NULL DEFAULT nextval('source_sid_seq'::regclass)
 sname   | character varying(255) |
 slink   | character varying(255) |
Индексы:
    "source_pkey" PRIMARY KEY, btree (sid)


\d year
                                Таблица "public.year"
 Столбец |          Тип           |                     Модификаторы
---------+------------------------+-----------------------------------------------------
 yid     | integer                | NOT NULL DEFAULT nextval('year_yid_seq'::regclass)
 yname   | character varying(255) |
Индексы:
    "year_pkey" PRIMARY KEY, btree (yid)


\d country
                                Таблица "public.country"
 Столбец |          Тип           |                      Модификаторы
---------+------------------------+-------------------------------------------------------
 cid     | integer                | NOT NULL DEFAULT nextval('country_cid_seq'::regclass)
 cname   | character varying(255) |
Индексы:
    "country_pkey" PRIMARY KEY, btree (cid)


\d value
                                Таблица "public.value"
 Столбец |          Тип           |                     Модификаторы
---------+------------------------+-----------------------------------------------------
 vid     | integer                | NOT NULL DEFAULT nextval('value_vid_seq'::regclass)
 vsid    | integer                |
 vyid    | integer                |
 vcid    | integer                |
 unit    | character varying(255) |
 val     | double precision       |
Индексы:
    "value_pkey" PRIMARY KEY, btree (vid)


             Таблица "public.tmp"
 Столбец |          Тип           | Модификаторы
---------+------------------------+--------------
 a0      | character varying(255) |
 a1      | character varying(255) |
 a2      | character varying(255) |
.............................................
 a39     | character varying(255) |
 a40     | character varying(255) |
 a41     | character varying(255) |
```

Ввод нормализованных данных

Источники

```
insert into source (sname, slink) values('ВВП, 1980–2018, ППС, млрд
$','http://svspb.net/danmark/vvp-stran.php');

select * from source;
 sid |            sname            |                 slink
-----+-----------------------------+---------------------------------------
   1 | ВВП, 1980–2018, ППС, млрд $ | http://svspb.net/danmark/vvp-stran.php
```

Страны, годы, значения

Вызов функций заполняет таблицу

```
select * from f_country();
select * from f_year();
select * from f_normalis();
```

Коды функций см. файл normalis.sql

ИРЧП
~~~~

Заливка данных в буферную таблицу

```
drop table tmp;
create table tmp (a0 varchar(255),
  a1 varchar(255), a2 varchar(255), a3 varchar(255),
  a4 varchar(255), a5 varchar(255), a6 varchar(255),
  a7 varchar(255), a8 varchar(255), a9 varchar(255),
  a10 varchar(255), a11 varchar(255), a12 varchar(255),
  a13 varchar(255), a14 varchar(255), a15 varchar(255),
  a16 varchar(255), a17 varchar(255), a18 varchar(255),
  a19 varchar(255), a20 varchar(255), a21 varchar(255),
  a22 varchar(255), a23 varchar(255), a24 varchar(255),
  a25 varchar(255), a26 varchar(255), a27 varchar(255),
  a28 varchar(255), a29 varchar(255), a30 varchar(255)
);

\copy tmp FROM '/home/an2k/py/richard/data/hdi.csv' DELIMITER ',' CSV
```

Согласование стран на русском и английском

```
create table country1 (cid1 serial primary key, cname1 varchar(255));
create table country_link (cid1 int, cid int);

\copy country_link FROM '/home/an2k/py/richard/data/country1-2.csv' DELIMITER ',' CSV

select country.cid, cname, cname1
  from country_link, country, country1
  where country_link.cid  = country.cid   and
        country_link.cid1 = country1.cid1
  order by country.cid;

alter table country add column cname_en varchar(255);

\d country
                              Таблица "public.country"
 Столбец  |          Тип           |                      Модификаторы
----------+------------------------+-------------------------------------------------------
 cid      | integer                | NOT NULL DEFAULT nextval('country_cid_seq'::regclass)
 cname    | character varying(255) |
 cname_en | character varying(255) |
Индексы:
    "country_pkey" PRIMARY KEY, btree (cid)


update country
  set cname_en = cname1
  from country1, country_link
  where country_link.cid  = country.cid and
        country_link.cid1 = country1.cid1
```

НИОКР
~~~~~

```
create table tmp3 (CountryISO3 varchar(255),CountryName varchar(255),IndicatorId
varchar(255),Indicator varchar(255),SubindicatorType varchar(255),a1960 varchar(255),a1961
varchar(255),a1962 varchar(255),a1963 varchar(255),a1964 varchar(255),a1965 varchar(255),a1966
varchar(255),a1967 varchar(255),a1968 varchar(255),a1969 varchar(255),a1970 varchar(255),a1971
varchar(255),a1972 varchar(255),a1973 varchar(255),a1974 varchar(255),a1975 varchar(255),a1976
varchar(255),a1977 varchar(255),a1978 varchar(255),a1979 varchar(255),a1980 varchar(255),a1981
varchar(255),a1982 varchar(255),a1983 varchar(255),a1984 varchar(255),a1985 varchar(255),a1986
varchar(255),a1987 varchar(255),a1988 varchar(255),a1989 varchar(255),a1990 varchar(255),a1991
varchar(255),a1992 varchar(255),a1993 varchar(255),a1994 varchar(255),a1995 varchar(255),a1996
varchar(255),a1997 varchar(255),a1998 varchar(255),a1999 varchar(255),a2000 varchar(255),a2001
varchar(255),a2002 varchar(255),a2003 varchar(255),a2004 varchar(255),a2005 varchar(255),a2006
varchar(255),a2007 varchar(255),a2008 varchar(255),a2009 varchar(255),a2010 varchar(255),a2011
varchar(255),a2012 varchar(255),a2013 varchar(255),a2014 varchar(255),a2015 varchar(255),a2016
varchar(255),a2017 varchar(255),a2018 varchar(255),a2019 varchar(255));
```

```
\copy tmp3 FROM '/home/an2k/py/richard/data/science4.csv' DELIMITER ',' CSV
```

ПУБЛИКАЦИИ В SCOPUS
~~~~~~~~~~~~~~~~~~~

https://www.scimagojr.com/countryrank.php

```
drop table tmp5;
create table tmp5 (Age  varchar(255),Rank varchar(255),Country varchar(255),Documents
varchar(255),Citable_documents varchar(255),Citations varchar(255),Self_citations
varchar(255),Citations_per_document varchar(255),H_index varchar(255));

\copy tmp5 FROM '/home/an2k/py/richard/data/public.csv' DELIMITER ';' CSV
```

SIPRI
~~~~~

Военные расходы, % от ВВП, SIPRI | https://www.sipri.org/databases/milex

```
create table tmp7 (country varchar(255), base varchar(255), notes varchar(255), a1949 varchar(255),
a1950 varchar(255), a1951 varchar(255), a1952 varchar(255), a1953 varchar(255), a1954 varchar(255),
a1955 varchar(255), a1956 varchar(255), a1957 varchar(255), a1958 varchar(255), a1959 varchar(255),
a1960 varchar(255), a1961 varchar(255), a1962 varchar(255), a1963 varchar(255), a1964 varchar(255),
a1965 varchar(255), a1966 varchar(255), a1967 varchar(255), a1968 varchar(255), a1969 varchar(255),
a1970 varchar(255), a1971 varchar(255), a1972 varchar(255), a1973 varchar(255), a1974 varchar(255),
a1975 varchar(255), a1976 varchar(255), a1977 varchar(255), a1978 varchar(255), a1979 varchar(255),
a1980 varchar(255), a1981 varchar(255), a1982 varchar(255), a1983 varchar(255), a1984 varchar(255),
a1985 varchar(255), a1986 varchar(255), a1987 varchar(255), a1988 varchar(255), a1989 varchar(255),
a1990 varchar(255), a1991 varchar(255), a1992 varchar(255), a1993 varchar(255), a1994 varchar(255),
a1995 varchar(255), a1996 varchar(255), a1997 varchar(255), a1998 varchar(255), a1999 varchar(255),
a2000 varchar(255), a2001 varchar(255), a2002 varchar(255), a2003 varchar(255), a2004 varchar(255),
a2005 varchar(255), a2006 varchar(255), a2007 varchar(255), a2008 varchar(255), a2009 varchar(255),
a2010 varchar(255), a2011 varchar(255), a2012 varchar(255), a2013 varchar(255), a2014 varchar(255),
a2015 varchar(255), a2016 varchar(255), a2017 varchar(255), a2018 varchar(255), a2018current
varchar(255));

\copy tmp7 FROM '/home/an2k/py/richard/data/sipriUSD2017.csv' DELIMITER ';' CSV
```

```
create table tmp9 (country varchar(255), notes varchar(255), a1949 varchar(255), a1950 varchar(255),
a1951 varchar(255), a1952 varchar(255), a1953 varchar(255), a1954 varchar(255), a1955 varchar(255),
a1956 varchar(255), a1957 varchar(255), a1958 varchar(255), a1959 varchar(255), a1960 varchar(255),
a1961 varchar(255), a1962 varchar(255), a1963 varchar(255), a1964 varchar(255), a1965 varchar(255),
a1966 varchar(255), a1967 varchar(255), a1968 varchar(255), a1969 varchar(255), a1970 varchar(255),
a1971 varchar(255), a1972 varchar(255), a1973 varchar(255), a1974 varchar(255), a1975 varchar(255),
a1976 varchar(255), a1977 varchar(255), a1978 varchar(255), a1979 varchar(255), a1980 varchar(255),
a1981 varchar(255), a1982 varchar(255), a1983 varchar(255), a1984 varchar(255), a1985 varchar(255),
a1986 varchar(255), a1987 varchar(255), a1988 varchar(255), a1989 varchar(255), a1990 varchar(255),
a1991 varchar(255), a1992 varchar(255), a1993 varchar(255), a1994 varchar(255), a1995 varchar(255),
a1996 varchar(255), a1997 varchar(255), a1998 varchar(255), a1999 varchar(255), a2000 varchar(255),
a2001 varchar(255), a2002 varchar(255), a2003 varchar(255), a2004 varchar(255), a2005 varchar(255),
a2006 varchar(255), a2007 varchar(255), a2008 varchar(255), a2009 varchar(255), a2010 varchar(255),
a2011 varchar(255), a2012 varchar(255), a2013 varchar(255), a2014 varchar(255), a2015 varchar(255),
a2016 varchar(255), a2017 varchar(255), a2018 varchar(255));

\copy tmp9 FROM '/home/an2k/py/richard/data/sipri_shareGDP.csv' DELIMITER ';' CSV
```

ПАТЕНТЫ
~~~~~~~

```
create table tmp10 (country varchar(255), code varchar(255), y1995 varchar(255), y1996 varchar(255),
y1997 varchar(255), y1998 varchar(255), y1999 varchar(255), y2000 varchar(255), y2001 varchar(255),
y2002 varchar(255), y2003 varchar(255), y2004 varchar(255), y2005 varchar(255), y2006 varchar(255),
y2007 varchar(255), y2008 varchar(255), y2009 varchar(255), y2010 varchar(255), y2011 varchar(255),
y2012 varchar(255), y2013 varchar(255), y2014 varchar(255), y2015 varchar(255), y2016 varchar(255),
y2017 varchar(255), y2018 varchar(255));
```
~~~~~~~

```
\copy tmp10 FROM '/home/an2k/py/richard/data/Patent.csv' DELIMITER ',' CSV

delete from tmp10 where
 y1995 is null and
 y1996 is null and
 y1997 is null and
 y1998 is null and
 y1999 is null and
 y2000 is null and
 y2001 is null and
 y2002 is null and
 y2003 is null and
 y2004 is null and
 y2005 is null and
 y2006 is null and
 y2007 is null and
 y2008 is null and
 y2009 is null and
 y2010 is null and
 y2011 is null and
 y2012 is null and
 y2013 is null and
 y2014 is null and
 y2015 is null and
 y2016 is null and
 y2017 is null and
 y2018 is null ;



delete from tmp10 where
 y1995 = '..' and
 y1996 = '..' and
 y1997 = '..' and
 y1998 = '..' and
 y1999 = '..' and
 y2000 = '..' and
 y2001 = '..' and
 y2002 = '..' and
 y2003 = '..' and
 y2004 = '..' and
 y2005 = '..' and
 y2006 = '..' and
 y2007 = '..' and
 y2008 = '..' and
 y2009 = '..' and
 y2010 = '..' and
 y2011 = '..' and
 y2012 = '..' and
 y2013 = '..' and
 y2014 = '..' and
 y2015 = '..' and
 y2016 = '..' and
 y2017 = '..' and
 y2018 = '..' ;



ЭЛЕКТРОЭНЕРГИЯ
~~~~~~~~~~~~~~

create table tmp11 (code varchar(255), year varchar(255), val varchar(255));

\copy tmp11 FROM '/home/an2k/py/richard/data/Electric.csv' DELIMITER ',' CSV


(Данных слишком мало)



4. PCA / ГЛАВНЫЕ КОМПОНЕНТЫ

drop table psa;
create table psa (
  pid serial primary key,
  pnc int,      -- номер PSA / главной компоненты
  pyid int,     -- id года
  pcid int,     -- id страны
  pval float    -- значение PSA / главной компоненты
```

);

После сбора и очистки данных для дальнейшей работы необходимо проанализировать их на предмет выделения их Главных компонент (PCA – principal component analysis).

Предварительный анализ показал, что для сохранения не менее 80% дисперсии собранных из 7 достоверных источников нормализованных данных достаточно 2-х компонент.

Скрипт, преобразующий собранную базу данных в две компоненты, использует библиотеки языка программирования Python.

Все файлы находятся в директории data в виде дерева:

```
data
├── dump
│   └── richard.sql     Дамп БД richard (СУБД Postgres (9.x)
├── inc
│   └── rch.py          Библиотека самописных функций (Python 3.x)
├── rich_pca.py         Скрипт расчета главных компонент (Python 3.x)
└── source              Директория источников данных
    ├── readmy.txt        Этот документ
    ├── vvp.zip           ВВП
    ├── sipri.zip         SIPRI
    ├── electr.zip        Выработка электроэнергии
    ├── ir4p.zip          ИРЧП
    ├── patent.zip        Патенты
    └── scopus.zip        SCOPUS
```

5. ВИЗУАЛИЗАЦИЯ ДАННЫХ

По существу это сайт на языке PHP с использованием СУБД Postgres.
Ниже приведена его структура

```
view
├── img                                     Картинки
│   └── formula_nomal.png
├── index.html                              Начальный HTML-файл
├── main.php                                Главный скрипт сайта (PHP 7.2.x)
├── lib.php                                 Самописная библиотека (PHP 7.2.x)
├── arr_color.php                           Цветовая дифференциация стран
├── pg_conn.php                             Параметры доступа к БД
├── gans.css                                Самописная таблица стилей (c)Gans
├── css                                     Таблица стилей (c)Bootstrap
│   ├── bootstrap.css
│   ├── bootstrap.css.map
│   ├── bootstrap-grid.css
│   ├── bootstrap-grid.css.map
│   ├── bootstrap-grid.min.css
│   ├── bootstrap-grid.min.css.map
│   ├── bootstrap.min.css
│   ├── bootstrap.min.css.map
│   ├── bootstrap-reboot.css
│   ├── bootstrap-reboot.css.map
│   ├── bootstrap-reboot.min.css
│   └── bootstrap-reboot.min.css.map
└── src -> /home/an2k/-Richard-/data/source  Ссылка на источники
```

6. РАЗНОЕ

Выгрузка таблицы стран из Postgres

```
copy (select * from country) to '/tmp/country.csv' delimiter ',' csv;
```