PD-R-Py 2019/2020

Praca domowa nr 4 (max. = 15 p.)

Maksymalna ocena: 15 p.

Termin oddania pracy: 16.05.2020, godz. 23:59

Do przesłania na adres prowadzącego laboratoria `M.Bartoszuk@mini.pw.edu.pl` lub `A.Cena@mini.pw.edu.pl` ze swojego konta pocztowego `*@*pw.edu.pl` **jeden** plik `.zip` (nie: `.rar`, `.7z` itp.) o nazwie `Nick_Nazwisko_Imie_NrAlbumu_pd4.zip`, zawierający:

- `Nick_Nazwisko_Imie_NrAlbumu_pd4.ipynb` (jeden raport)
- `Nick_Nazwisko_Imie_NrAlbumu_pd4.hmtl` (ściągnięta wersja powyższego w formacie `.html` – zob. `File -> Downlowad as -> html` w notatniku `Jupyter`).
- zdjęcie lub skan oświadczenia o samodzielności wykonanej pracy (treść znajduje się w sekcji `Oświadczenie`, nazwa pliku `Nazwisko_Imie_nrAlbumu`)

Uwaga: temat wiadomości to `[PDRPy] Praca domowa nr 4`.

> `Nick` to wymyślony przez Ciebie identyfikator, który pojawi się w arkuszu ocen i zapewni Ci odpowiednią anonimowość. Zapamiętaj go, bo przesyłając kolejne prace domowe, będziesz używała/używał tego samego nicka.

1 Zbiory danych

Będziemy znów pracować na uproszczonym zrzucie zanonimizowanych danych z serwisu https://travel.stackexchange.com/, który składa się z następujących ramek danych:

- Badges.csv.gz
- Comments.csv.gz
- PostLinks.csv.gz
- Posts.csv.gz
- Tags.csv.gz
- Users.csv.gz
- Votes.csv.gz

Dane pobrać można także ze strony:

- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Badges.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Comments.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/PostLinks.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Posts.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Tags.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Users.csv.gz
- http://www.gagolewski.com/resources/data/travel_stackexchange_com/Votes.csv.gz

Przed przystąpieniem do rozwiązywania zadań zapoznaj się z ww. serwisem oraz znaczeniem poszczególnych kolumn w ww. tabelach, zob. http://www.gagolewski.com/resources/data/travel_stackexchange_com/readme.txt .

Każdą z ramek danych należy wyeksportować do bazy danych SQLite przy użyciu wywołania metody `to_sql()` w klasie `pandas.DataFrame`.

Przykładowe wywołanie – ładowanie zbioru `Tags`:

```
options(stringsAsFactors=FALSE)
# ww. pliki pobralismy do katalogu pd1/travel_stackexchange_com
Tags <- pd.read_csv("pd1/travel_stackexchange_com/Tags.csv.gz",
                    compression = "gzip")
```

2 Informacje ogólne

Rozwiąż poniższe zadania przy użyciu wywołań funkcji i metod z pakietu `pandas`. Każdemu z 7 poleceń SQL powinny odpowiadać dwa równoważne sposoby ich implementacji, kolejno:

1. wywołanie `pandas.read_sql_query("""zapytanie SQL""")`;
2. wywołanie ciągu „zwykłych” metod i funkcji z pakietu `pandas`.

Upewnij się, że zwracane wyniki są ze sobą tożsame (ewentualnie z dokładnością do permutacji wierszy wynikowych ramek danych – przydatna może być metoda `equals()`).

W szczególności należy zagwarantować, że w każdym przypadku wynik jest klasy `DataFrame` a nie `Series`.

Wszystkie rozwiązania umieść w jednym (estetycznie sformatowanym) raporcie Jupyter.

3 Zadania do rozwiązania

```
--- 1)
SELECT
    Posts.Title,
    UpVotesPerYear.Year,
    MAX(UpVotesPerYear.Count) AS Count
FROM (
        SELECT
            PostId,
            COUNT(*) AS Count,
            STRFTIME('%Y', Votes.CreationDate) AS Year
        FROM Votes
        WHERE VoteTypeId=2
        GROUP BY PostId, Year
    ) AS UpVotesPerYear
JOIN Posts ON Posts.Id=UpVotesPerYear.PostId
WHERE Posts.PostTypeId=1
GROUP BY Year
```

```sql
--- 3)
SELECT
    Posts.ID,
    Posts.Title,
    Posts2.PositiveAnswerCount
FROM Posts
JOIN (
        SELECT
            Posts.ParentID,
            COUNT(*) AS PositiveAnswerCount
        FROM Posts
        WHERE Posts.PostTypeID=2 AND Posts.Score>0
        GROUP BY Posts.ParentID
    ) AS Posts2
    ON Posts.ID=Posts2.ParentID
ORDER BY Posts2.PositiveAnswerCount DESC
LIMIT 10
```

```sql
--- 6)
SELECT DISTINCT
    Users.Id,
    Users.DisplayName,
    Users.Reputation,
    Users.Age,
    Users.Location
FROM (
        SELECT
            Name, UserID
        FROM Badges
        WHERE Name IN (
            SELECT
                Name
            FROM Badges
            WHERE Class=1
            GROUP BY Name
            HAVING COUNT(*) BETWEEN 2 AND 10
        )
        AND Class=1
    ) AS ValuableBadges
JOIN Users ON ValuableBadges.UserId=Users.Id
```

4 Oświadczenie

Do przesłanej pracy należy dołączyć własnoręcznie podpisane oświadczenie o następującej treści:

Oświadcznie w sprawie pracy projektowej nr 4

Oświadczam, że niniejsza praca stanowiąca podstawę do uznania efektów uczenia się z przedmiotu

Przetwarzanie danych w języku R i Python

została wykonana przeze mnie samodzielnie.

Imię i Nazwisko

Nr albumu