PDU 2018/2019

Praca domowa nr 3 (max. = 30 p.)

Zadanie rozwiązujemy w grupach dwuosobowych. Propozycje składu grup zgłaszamy prowadzącemu laboratoria. Można używać zarówno R jak i Pythona.

Termin oddania pracy: 03.06.2019, godz. 23:59.

Do przesłania na adres `cena@rexamine.com` lub `zogala@rexamine.com` (swojego prowadzącego laboratoria) – **jedno archiwum `.zip`**[1] o nazwie typu `Nick1_Nazwisko1_Nick2_Nazwisko2_pd3.zip`, w którym znajdziemy:

- prezentację (slajdy) zawierającą omówienie sposobu rozwiązania zadania oraz przedstawiającą wyniki analizy danych (PDF lub HTML) – to *głównie* na jej podstawie zostanie wystawiona ocena;
- wszystkie skrypty/notatniki/raporty pozwalające na odtworzenie zawartych w prezentacji wyników;
- dane pośrednie, na podstawie których zostały wygenerowane ostateczne wyniki (pliki `.csv`, `.json`, `.xml` itp.); uwaga: *nie* dodajemy plików zawierających dane surowe – przesyłany plik `.zip` powinien być „rozsądnych" rozmiarów.

Nazwy plików nie powinny zawierać polskich liter diakrytyzowanych (przekształć $ą \to a$ itd.). Treść wysyłanego e-maila nie może być pusta. Uwaga: tytuł wiadomości to `[PDU] Praca domowa nr 3`. Użyj koniecznie tego samego `Nick`a, co w pracy domowej nr 1.

Prezentacje: Na XIV i XV zajęciach laboratoryjnych każda dwuosobowa grupa przedstawi najciekawsze ich zdaniem wyniki (10 minut na projekt + 5 minut na dyskusję i pytania od słuchaczy; polecamy przyjście z własnym laptopem). Wygłoszenie prezentacji jest warunkiem koniecznym uzyskania pozytywnej oceny.

1 Dane do analizy

Podczas poprzedniej pracy domowej pracowaliśmy na uproszczonych danych z forum Travel Stack Exchange – pora wypłynąć na (jeszcze) szersze wody!

Na stronie https://archive.org/details/stackexchange mamy dostępne zanonimizowane zrzuty ze wszystkich serwisów Stack Exchange. We wszystkich przypadkach (bodaj oprócz giganta StackOverflow) każdy serwis zapisany jest w postaci jednego archiwum .7z, które zawiera 8 tabel (plików XML[2]); ich opis znajdziemy na stronie https://archive.org/27/items/stackexchange/readme.txt oraz https://meta.stackexchange.com/questions/2677 .

Należy wybrać co najmniej trzy serwisy do analizy, w tym jeden z nich musi być *niemały* (>100 MB). Niniejsza praca domowa to prawdziwe wyzwanie data science – to każda grupa sama stawia ciekawe (dla siebie i słuchaczy) pytania i generuje na nie odpowiedzi.

Interesują nas zagadnienia dotyczące konkretnych serwisów, ale i porównania między serwisami. Stan „na dziś" i trendy w czasie. Rzeczy popularne i rarytasy. Różnice i podobieństwa. You name it.

2 Ocena

Ocenę co najmniej dostateczną (> 50%) uzyskają prace, które spełniają następujące kryteria:

1. zawierają kod potrzebny do załadowania zbiorów danych,

[1] A więc nie: '.rar', '.7z' itp.

[2] 'Badges', 'Comments', 'PostHistory', 'PostLinks', 'Posts', 'Tags', 'Users' oraz 'Votes'

2. stworzą kod, dzięki któremu zostaną wygenerowane co najmniej trzy ciekawe wyniki (odpowiedzi na pytania „badawcze” w postaci wykresów/tabel/itp.),
3. przedstawią uzyskane wyniki w formie prezentacji.

Każda dodatkowa analiza czy nietrywialna zastosowana technika będzie wpływać pozytywnie na ocenę (np. wykresy interaktywne, animacje, aplikacje webowe, mapy, algorytmy i struktury danych umożliwiające poprawę szybkości wykonywanych analiz, własne implementacje metod znanych z literatury (z autorskimi modyfikacjami) itp.). W szczególności, ocenę maksymalną (bardzo dobrą) uzyskają tylko prace naprawdę wyróżniające się pod względem jakościowymi i merytorycznym.