

PDU 2020/2021

Praca domowa nr 3 (max. = 40 p.)

Maksymalna ocena: 40 p.

Termin oddania pracy: 08.06.2021, godz. 23:59

Zadanie rozwiążemy w grupach dwuosobowych lub trzyosobowych.

Prace należy przesłać za pośrednictwem platformy Moodle – **jedno archiwum .zip¹** o nazwie typu Nazwisko1_NrAlbumu1_Nazwisko2_NrAlbumu2_pd3.zip, w którym znajdziemy:

- prezentację (slajdy) zawierającą omówienie sposobu rozwiązania zadania oraz przedstawiającą wyniki analizy danych (PDF lub HTML)
- plik `link.txt` zawierający link do filmu na youtube z wygłoszoną prezentacją – to *głównie* na podstawie prezentacji i filmu zostanie wystawiona ocena;
- wszystkie skrypty/moduły pozwalające na odtworzenie zawartych w prezentacji wyników;
- dane pośrednie, na podstawie których zostały wygenerowane ostateczne wyniki (pliki `.csv`, `.json`, `.xml` itp.) - jeśli takie były wytworzone; uwaga: *nie* dodajemy plików zawierających dane surowe – przesyłany plik `.zip` powinien być „rozsądnym” rozmiarów.

Nazwy plików nie powinny zawierać polskich liter diakrytyzowanych (przekształć $q \rightarrow a$ itd.).

Prezentacje: Na XIV i XV zajęciach laboratoryjnych w trybie “sprzed 2020 r” każda grupa przedstawiłaby swoje wyniki w formie prezentacji (10 minut na projekt + 5 minut na dyskusję i pytania od słuchaczy). Ze względu na tryb pracy zdalnej, prezentacje należy przygotować w formie filmu udostępnionego na youtube.

Film powinien być nagrany i zmontowany w ramach zespołu: * np. każdy uczestnik nagrywa część dotyczącą swojego wystąpienia lub nagrana jest prezentacja z wykorzystaniem MS Teams, * itp. * film udostępniają Państwo w ramach kanału youtube jako **unlisted** (wtedy będzie dostępny tylko dla osób posiadających link), * wystarczy gdy film będzie zawierał rzut ekranu z prezentacją.

Film powinien trwać max. 10 minut w przypadku grupy dwuosobowej, 15 min w przypadku grupy trzyosobowej.

Link do filmu zostanie udostępniony do obejrzenia dla pozostałych uczestników kursu.

Wygłoszenie prezentacji czyli przygotowanie filmu jest warunkiem koniecznym uzyskania pozytywnej oceny.

1 Dane do analizy

Będziemy kontynuować pracę z danymi udostępnionymi przez sieć Stack Exchange, jednak nie ograniczymy się tylko do danych z forum Travel Stack Exchange – tym razem mają Państwo swobodę wyboru interesujących Państwa portali.

Na stronie <https://archive.org/details/stackexchange> mamy dostępne zanonimizowane zrzuty ze wszystkich serwisów Stack Exchange. We wszystkich przypadkach (z wyjątkiem StackOverflow) każdy serwis zapisany jest

¹A więc nie: ‘rar’, ‘7z’ itp.

w postaci jednego archiwum .7z, które zawiera 8 tabel (plików XML²); 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 *niemalą* (>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.

2 Ocena

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

1. zawierają kod potrzebny do wczytania/załadowania zbiorów danych (z dowolnego forum),
2. stworzą kod, dzięki któremu zostaną wygenerowane co najmniej **trzy/cztery**★ ciekawe wyniki (odpowiedzi na pytania „badawcze” w postaci wykresów/tabel/itp.),
3. przedstawiają uzyskane wyniki w formie prezentacji/filmu.

★ - trzy dla zespołów dwuosobowych, cztery dla trzyosobowych.

Każda dodatkowa analiza czy nietrywialna zastosowana technika będzie wpływać pozytywnie na ocenę – np. wykorzystanie GITHUB do pracy zespołowej (należy wtedy załączyć link do repozytorium i dodać prowadzących), 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.

²‘Badges’, ‘Comments’, ‘PostHistory’, ‘PostLinks’, ‘Posts’, ‘Tags’, ‘Users’ oraz ‘Votes’