

ZTB, Projekty 2024

Igor Wojnicki

2024-10-18

Spis treści

1	Warunki ogólne	1
2	Zasady oceniania	2
3	Temat	2
4	Cele	3
5	Kroki do wykonania	4
6	Zdefiniowanie parametrów oceny rozwiązań	4
7	Harmonogram	4
8	Lista kontrolna dokumentacji	5

1 Warunki ogólne

1. Grupy. 2-osobowe.
2. Temat. Każda grupa pracuje niezależnie nad tym samym tematem.
3. Środowisko pracy. Każda grupa otrzyma oddzielne repozytorium GitLab, które powinno służyć do zarządzania projektem (tablica/problemy), dokumentowania go oraz przechowywania kodu i innych danych. Wszystkie odpowiednie pliki powinny znajdować się w gałęzi *main*. Przed pobraniem repozytorium projektu **trzeba** wejść na stronę <https://gitlab.kis.agh.edu.pl> i zalogować się za pomocą SSO AGH.
4. Wybór technologii. O ile nie zaznaczono inaczej, wybrane technologie i architektura powinny być jak najlżejsze, np.: wymagają minimalnej konfiguracji, łatwe do uruchomienia i oceny. Projekty muszą działać pod Linuksem. Wybór architektury musi zostać zatwierdzony przez instruktora.
5. Samowystarczalność. Projekty powinny być samowystarczalne, m.in. zawierać kod potrzebny do skonfigurowania zależności, utworzenia struktur baz danych i wypełnienia ich danymi. Dane źródłowe należy pobrać ze źródeł zewnętrznych w sposób automatyczny lub uwzględnić w repozytorium.

6. Dokumentacja. Projekty powinny zawierać dokumentację w pliku `README.md` lub `README.org` w katalogu głównym gałęzi *main*, odpowiednio w postaci plików Markdown lub org-mode. Dokumentacja musi zawierać instrukcje dotyczące wdrożenia i uruchomienia oprogramowania, określić wymagania wstępne oraz udokumentować proces projektowania i wdrażania projektu. Musi zawierać również informacje o roli poszczególnych osób w projekcie i opis, kto co zrobił. Szczegółowe wymagania odnośnie dokumentacji zamieszczone są w sekcji Lista kontrolna dokumentacji.
7. Wyniki. Wyniki powinny być łatwe do oceny i powtarzalne. Należy dostarczyć instrukcję krok po kroku, jak odtworzyć wyniki. Szczegółowe wymagania odnośnie dokumentacji zamieszczone są w sekcji Lista kontrolna dokumentacji.
8. Samoocena. Samoocena projektów jest obowiązkowa. Należy omówić efektywność oraz proponowane strategie mitygacji zidentyfikowanych niedociągnięć w przyszłości. Szczegółowe wymagania odnośnie dokumentacji zamieszczone są w sekcji Lista kontrolna dokumentacji.

2 Zasady oceniania

Projekty będą oceniane na podstawie pkt 4-8 Warunków ogólnych oraz pod względem technicznym. Szczegółowe wymagania odnośnie dokumentacji zamieszczone są w sekcji Lista kontrolna dokumentacji.

Każda grupa musi skonsultować kamienie milowe projektu na spotkaniu, jak wskazano w sekcji Kroki do wykonania.

Kamień milowy 1 musi być skonsultowany przed zdefiniowaniem parametrów oceny rozwiązań (patrz sekcja Harmonogram), pod rygorem oceny niedostatecznej bez możliwości poprawy.

Do zaliczenia konieczna jest prezentacja na zajęciach. Aby zaliczyć zajęcia projektowe w terminie, należy to zrobić najpóźniej na tydzień przed zakończeniem semestru. Komplet oprogramowania i dokumentacji musi być umieszczony w repozytorium przed prezentacją.

W przypadku otrzymania oceny negatywnej projekt można odrabiać dwukrotnie po zakończeniu semestru. Informacje o poprawionej wersji dokumentacji lub oprogramowania należy przesłać e-mailem co najmniej na tydzień przed zakończeniem sesji poprawkowej.

ocena	%
5	90
4,5	80
4	70
3,5	60
3	50
2	

3 Temat

Dany jest plik wartości rozdzielanych przecinkami (CSV) `taxonomy_iw.csv.gz`, który tworzy graf skierowany. Reprezentuje kategorie klasyfikacji głównego tematu Wikipedii. Każda linia to pojedynczy rekord z dwoma polami wskazującymi relację kategoria-podkategoria, np. :

```
"1880s_films", "1889_films"
```

wskazuje, że istnieje kategoria `"1880s_films"`, która zawiera podkategorię `"1889_films"`. W pliku znajduje się 5771611 rekordów.

Dany jest również plik w formacie CSV o nazwie `popularity_iw.csv.gz` zawierający informacje o popularności kategorii. Każda linia to pojedynczy rekord z dwoma polami określającymi nazwę węzła oraz wartość całkowitą reprezentującą jego popularność, np.:

```
"1889_films",34
```

wskazuje, że `"1889_films"` ma popularność 34. W pliku znajduje się 952453 rekordów.

W ramach projektu należy:

1. zidentyfikować i wybrać odpowiednią technologię bazodanową (uwaga: nie może to być Neo4J), technologia musi być otwarta i udostępniać kod źródłowy na otwartej licencji,
2. zaprojektować bazę danych,
3. załadować ją powyższymi danymi,
4. opracować i przetestować narzędzie wiersza poleceń, które ma spełniać cele opisane poniżej.

Narzędzie musi nosić nazwę `dbcli` i powinno przyjmować numer celu oraz opcjonalne argumenty. Przykładowo, aby uruchomić go dla celu numer 1 z dodatkowym parametrem wskazującym węzeł `"1880s_films"`, należy go uruchomić jako:

```
dbcli 1 "1880s_films"
```

4 Cele

Narzędzie wiersza poleceń musi spełniać następujące cele:

1. znajduje wszystkie dzieci danego węzła,
2. zlicza wszystkie dzieci danego węzła,
3. znajduje wszystkie wnuki danego węzła,
4. znajduje wszystkich rodziców danego węzła,
5. zlicza wszystkich rodziców danego węzła,
6. znajduje wszystkich dziadków danego węzła,
7. liczy, ile jest węzłów o unikatowych nazwach,
8. znajduje węzły, które nie są podkategorią żadnego innego węzła,
9. zlicza węzły z celu 8,
10. znajduje węzły z największą liczbą dzieci, może być ich więcej,
11. znajduje węzły z najmniejszą liczbą dzieci (liczba dzieci jest większa od zera),
12. zmienia nazwę danego węzła,
13. zmienia popularność danego węzła,
14. znajduje wszystkie ścieżki pomiędzy dwoma podanymi węzłami, z krawędziami skierowanymi od pierwszego do drugiego węzła.

15. zlicza węzły z celu 14,
16. policzy popularność w sąsiedztwie węzła o zadany promieniu; parametrami są: nazwa węzła oraz promień sąsiedztwa; popularność sąsiedztwa jest sumą popularności danego węzła oraz wszystkich węzłów należących do sąsiedztwa,
17. policzy popularność na najkrótszej ścieżce między dwoma danymi węzłami, zgodnie ze skierowaniem; popularność na najkrótszej ścieżce jest sumą popularności wszystkich węzłów znajdujących się na najkrótszej ścieżce,
18. znajdzie skierowaną ścieżkę pomiędzy dwoma węzłami o największej popularności spośród wszystkich ścieżek pomiędzy tymi węzłami.

5 Kroki do wykonania

Uwaga: kamienie milowe muszą być skonsultowane na zajęciach; nie można skonsultować więcej niż 1 kamienia milowego na raz.

1. Znajdź i wybierz odpowiednią bazę danych. Zaprojektuj bazę danych. **Skonsultuj** swój wybór i projekt. (**Kamień milowy 1**)
2. Zaimportuj plik CSV do bazy danych. **Przedstaw** narzędzia i sposób przeprowadzania importu. (**Kamień milowy 2**)
3. Zaprojektuj narzędzie dbcli. **Skonsultuj** szczegóły działania narzędzia, w szczególności zapytania do bazy danych. (**Kamień milowy 3**)
4. Zaimplementuj potrzebne komponenty, zoptymalizuj wydajność, użyj indeksowania, jeśli to konieczne.
5. Zaprezentuj narzędzie. (**Kamień milowy 4**)

6 Zdefiniowanie parametrów oceny rozwiązań

Na spotkaniu projektowym wszystkie grupy muszą określić za pomocą jakich parametrów można ocenić powstające implementacje. W szczególności:

- jakie wielkości powinny być mierzone,
- których zapytań użyć,
- jakie mają być konkretne wartości parametrów zapytań,
- czy oprócz zapytań mierzyć coś innego?

7 Harmonogram

1. Spotkanie wprowadzające,
2. Konsultacje **kamienia milowego 1**,
3. Konsultacje **kamienia milowego 1**,

4. Zdefiniowanie parametrów oceny rozwiązań,
5. Konsultacje **kamienia milowego 2**
6. Konsultacje **kamienia milowego 3**
7. Prezentacja, **kamień milowy 4**

8 Lista kontrolna dokumentacji

Upewnij się, że dokumentacja składa się z poniższych elementów.

1. ☐ Wybór technologii.
2. ☐ Architektura: komponenty i interakcje, schemat (opcjonalny).
3. ☐ Wymagania i zależności (moduły oprogramowania, bazy danych itp.).
4. ☐ Instrukcje instalacji i konfiguracji.
5. ☐ Instrukcja obsługi (jak uruchomić program).
6. ☐ Proces projektowania i wdrażania krok po kroku.
7. ☐ Role wszystkich osób w projekcie i opis tego, kto co zrobił.
8. ☐ Wyniki.
9. ☐ Instrukcja krok po kroku jak odtworzyć wyniki.
10. ☐ Samoocena: należy omówić efektywność.
11. ☐ Strategie przyszłego łagodzenia zidentyfikowanych niedociągnięć.