SPDB – projekty, semestr 2020Z

Procedura wyboru tematu projektu do realizacji:

1. Należy przesłać e-mailem preferencje (minimum 3 tematy) do prowadzących (r.bembenik@ii.pw.edu.pl, g.protaziuk@elka.pw.edu.pl) do 03.11.2020 r. włącznie (zapisy na temat będą realizowane wg kolejności zgłoszeń i preferencji) – z wykorzystaniem konta/adresu politechnicznego.
2. Osoby zgłaszające grupę powinny wysłać ten sam e-mail do wiadomości wszystkich pozostałych członków zespołu.
3. Osoby, które chcą realizować własny temat muszą uzgodnić cel i zakres projektu z jednym z prowadzących do dnia 03.11.2020 r.
4. Osoby, które nie prześlą preferencji w wyznaczonym terminie zostaną przypisane do tematu projektu arbitralnie wybranego przez prowadzących.

Prezentacja kluczowych elementów rozwiązania:

Kluczowe elementy rozwiązania należy przedstawić osobiście do 11.12.2020 r. Przedstawienie obejmuje:

- przedstawienie idei rozwiązania w formie prezentacji (np. w formacie .ppt);
- omówienie planowanego rozwiązania.

Przed prezentacją kluczowych elementów projektu należy uzgodnić z prowadzącym dokładny termin.

Oddanie projektu:

Projekt należy oddać osobiście do 19.01.2021 r. z zachowaniem możliwości zwolnienia z jednego ze sprawdzianów lub do 27.01.2021 r., bez możliwości zwolnienia ze sprawdzianu.

Oddanie projektu obejmuje:

1. prezentację pokazującą główne zagadnienia związane z realizowanym projektem – należy przygotować prezentację (np. w formacie .ppt);
2. pokaz działania oprogramowania (jeśli jest to przedmiotem projektu);
3. rozmowę dotyczącą uzyskanych wyników i wniosków.

Przed oddaniem projektu należy uzgodnić z prowadzącym dokładny termin oddania oraz przesłać odpowiednio wcześniej (absolutne minimum to jeden dzień roboczy przed oddaniem) dokumentację projektu.

Tematy projektów

1. Analiza danych dotyczących przejazdów hulajnogami elektrycznymi w Warszawie (lub dodatkowo w innym polskim mieście/miastach). Projekt polega na zgromadzeniu danych dotyczących przejazdów hulajnogami od kilku dostawców (np. Lime, Hive, Bolt) w okresie min. 1 miesiąca i dokonaniu analizy tych danych. W szczególności zadanie obejmuje:

- stworzenie aplikacji do pobierania danych i pobranie potrzebnych danych,
- analizę przemierzanych prze hulajnogi tras w dni robocze vs. dni świąteczne/weekendy,
- znalezienie grup miejsc rozpoczynania podróży i miejsc docelowych,
- wizualizację wyników z wykorzystaniem R, PowerBI lub Tableau.

Dokumentacja projektowa powinna zawierać:

- zbiór danych i opis danych wykorzystanych w eksperymentach;
- wyniki przeprowadzonych analiz i wnioski dotyczące możliwości ich praktycznego wykorzystania;

- charakterystykę stworzonych wizualizacji.

Proponowana liczba osób w grupie: 2-3.

2. Analiza popularności POI w Warszawie (lub dodatkowo w innym polskim mieście/miastach). Projekt polega na zgromadzeniu danych dotyczących zameldowań użytkowników używających serwisów takich jak Twitter, Foursquare, Swarmapp, Facebook w okresie min. 1 miesiąca. Podczas pobierania i analizy danych należy oprzeć się na przykładach z artykułów podanych w sekcji **Wykorzystanie danych z sieci społecznościowych LBSN** tego dokumentu.

Zadanie obejmuje:
- stworzenie aplikacji do pobierania danych i pobranie potrzebnych danych,
- analizę najbardziej popularnych (najczęściej odwiedzanych) miejsc,
- analizę najpopularniejszych sekwencji miejsc odwiedzanych przez użytkowników,
- wizualizację wyników z wykorzystaniem R, PowerBI lub Tableau.

Dokumentacja projektowa powinna zawierać:
- zbiór danych i opis danych wykorzystanych w eksperymentach;
- wyniki przeprowadzonych analiz i wnioski dotyczące możliwości ich praktycznego wykorzystania;
- charakterystykę stworzonych wizualizacji.

Proponowana liczba osób w grupie: 2-3.

3. Znajdowanie najlepszej trasy. Implementacja aplikacji do wyznaczania najlepszej trasy wg zadanych parametrów, opracowanie metody oraz wykonanie testów sprawdzających praktyczną użyteczność zaproponowanych rozwiązań (wyznaczanie trasy w miastach dla kilku lub więcej podanych miejsc do odwiedzenia). Podstawowymi parametrami do wyboru najlepszej trasy jest: czas i odległość. Dodatkowym wymogiem jest minimalizacja skrętów w lewo.

Proponowana liczba osób w grupie: 2.
Dokumentacja końcowa powinna zawierać:
- Opis metody wyznaczania najlepszej drogi, w tym sposobu wyboru najlepszej trasy. Opis modelu danych.
- Opis architektury aplikacji.
- Informacje o implementacji (wykorzystane algorytmy + skomentowany kod źródłowy).
- Opis i wyniki przeprowadzonych testów.

4. Implementacja aplikacji pozwalającej na wyznaczenie optymalnej trasy turystycznej (możliwej do przebycia pieszo, samochodem, komunikacją miejską, lub kombinacją tych sposobów) w celu odwiedzenia zadanych miejsc (miejscami są podane przez użytkownika obiekty turystyczne i/lub restauracje na trasie wycieczki). Dla każdego miejsca użytkownik określa zakładany czas pobytu (może to być przedział np. do 2 do 3 godzin). Każde miejsce ma podane godziny otwarcia. Dodatkowo użytkownik podaje: miejsce i godzinę rozpoczęcia wycieczki, może podać: godzinę zakończenia, oraz że chce odwiedzić dane miejsce (ew. typ miejsca) w określonym przedziale czasu. Wynikiem działania, oprócz wyznaczenia trasy, powinna być informacja o czasie potrzebnym na realizację wycieczki. Aplikacja powinna umożliwiać wizualizację trasy na mapie.

Proponowana liczba osób w grupie: 2-3
Dokumentacja końcowa powinna zawierać:

- Wyniki eksperymentów pozwalające na ocenę jakości wyników.
- Informacje o implementacji (skomentowany kod źródłowy, zastosowany model danych, architektura aplikacji).

5. Porównanie sposobów indeksowania danych przestrzennych w wybranych systemach baz NoSQL (np. MongoDB, Casandra). W ramach realizacji projektu należy opracować zestawienie zawierające:
   - informacje nt. dostępnych indeksów dla danych przestrzennych,
   - dla danych o różnej wielkości:
     - rozmiar i czas tworzenia indeksów,
     - czas wykonania przykładowych zapytań, w których wykorzystany jest indeks przestrzenny (dla każdej z baz danych zapytania powinny zwracać ten sam wynik).

   Do testowania należy wykorzystać dane udostępnione przez serwisy oferujące dane przestrzenne (np. serwis OpenStreetMap http://www.openstreetmap.org/).

   Proponowana liczba osób w grupie: 2 (jedna osoba na SZBD)
   Dokumentacja końcowa powinna zawierać:
   - Schemat danych wraz z opisem.
   - Informacje o zastosowanych indeksach przestrzennych i nieprzestrzennych.
   - Informacje o danych: krótka charakterystyka (ile obiektów, jakich typów, etc.).
   - Skrypty wykorzystywane w trakcie testów, wyniki testów oraz komentarz do testów.

6. Porównanie wydajności modułów przestrzennych w systemach baz danych (MySQL, SQL Server, PostgreSQL, Oracle, MongoDB, Casandra,…) W ramach realizacji projektu należy przetestować czas realizacji zapytań typu:
   - punktowe: podać wszystkie obiekty, które zawierają dany punkt;
   - przecięcie: znaleźć obiekty, które mają część wspólną z danym odcinkiem/innym obiektem;
   - zawieranie: podać wszystkie obiekty zawarte w danym obszarze (obszar: prostokąt lub większy obiekt);
   - zapytania dotyczące sąsiadów: wyznaczanie k najbliższych sąsiadów, podać wszystkie obiekty (określonego typu) znajdujące się w zadanej odległości;
   - obliczanie pola obiektów;
   - obliczanie odległości,
   - zapytań skorelowanych.

   oraz skalowalność systemu względem ilości danych.

   Do testowania należy wykorzystać dane udostępnione przez serwisy oferujące dane przestrzenne (np. serwis OpenStreetMap http://www.openstreetmap.org/).

   Proponowana liczba osób w grupie: jeden SZBD na osobę.
   Dokumentacja końcowa powinna zawierać:
   - Schemat danych wraz z opisem.
   - Informacje o zastosowanych indeksach przestrzennych i nieprzestrzennych.
   - Informacje o danych: krótka charakterystyka (rozmiar danych, ile obiektów, jakich typów, etc.).
   - Zapytania wykorzystane w testach – skrypt do wykonania w SZBD

- Plan wykonania zapytań – przynajmniej jeden na dany typ zapylania.
- Wyniki testów oraz komentarz do testów.

7. Implementacja wybranego algorytmu eksploracji danych.
   Algorytmy, które można badać/implementować są opisane w artykułach, których listę zamieszczono na końcu bieżącego dokumentu. Są to przykładowe algorytmy. Możliwe jest wybranie innego algorytmu (po wcześniejszym uzgodnieniu).

   Proponowana liczba osób w grupie: 1
   Dokumentacja końcowa powinna zawierać:
   - Opis zastosowanego modelu
   - Opis (pseudokod) algorytmu, jeżeli były dokonywane zmiany w porównaniu z algorytmem oryginalnym.
   - Wyniki testów (wykresy + komentarze).
   - Skomentowany kod.

8. Zbadanie funkcjonalności bibliotek R *sf* i *sp* dotyczącej metod reprezentacji i wizualizacji wektorowych danych przestrzennych.
   Zadanie obejmuje:
   - opis dostępnych reprezentacji i metod wizualizacji danych wektorowych;
   - zbadanie przydatności tych metod do analizy danych przestrzennych (poprawność działania, czas wykonania dla różnych wielkości danych, łatwość użycia funkcji, szczegółowość dostępnych opisów funkcji – pomocy, ocena użytkownika);
   - dobór i transformacje układu współrzędnych;
   - transformacje danych pochodzących z innych źródeł.

   Dokumentacja projektowa powinna zawierać:
   - opis metod reprezentacji i wizualizacji;
   - opis środowiska, w którym zostały wykonane testy;
   - zbiory danych i opis danych wykorzystanych w eksperymentach;
   - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
   - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

   Proponowana liczba osób w grupie: 2.

9. Porównanie metod reprezentacji i wizualizacji danych przestrzennych dostępnych w bibliotekach R: *spatstat* (ppp) oraz *sp*.
   Zadanie obejmuje:
   - opis dostępnych reprezentacji i metod wizualizacji danych przestrzennych;
   - porównanie metod wizualizacji (poprawność działania, czas wykonania dla różnych wielkości danych, łatwość użycia funkcji, szczegółowość i kompletność dostępnych opisów funkcji, ocena użytkownika);
   - dobór i transformacje układu współrzędnych;
   - transformacje danych pochodzących z innych źródeł.

   Dokumentacja projektowa powinna zawierać:
   - opis metod reprezentacji i wizualizacji;
   - opis środowiska, w którym zostały wykonane testy;

- zbiory danych i opis danych wykorzystanych w eksperymentach;
- wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
- skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

Proponowana liczba osób w grupie: 2.

10. Zbadania możliwości realizacji zapytań o dane przestrzenne w środowisku R.
    Celem projektu jest analiza dostępnych w środowisku R funkcji do wyszukiwania danych przestrzennych w zbiorach. W zadaniu należy wskazać dostępne możliwości wyszukiwania: na podstawie relacji topologicznych, relacji odległościowych, możliwości złączeń zbiorów danych przestrzennych, funkcji agregujących.

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;
    - zbiory danych i opis danych wykorzystanych w eksperymentach;
    - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
    - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

    Proponowana liczba osób w grupie: 1.

11. Zbadania możliwości przekształceń danych przestrzennych w środowisku R.
    Celem projektu jest analiza dostępnych w środowisku R funkcji do przekształcania danych przestrzennych. W zadaniu należy zbadać dostępne funkcje dotyczące: upraszczania danych, wyznaczania środków oraz buforów, transformacji afiniczne, przycinania obiektów, tworzenia unii geometrycznych oraz metod transformacji typów.

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;
    - zbiory danych i opis danych wykorzystanych w eksperymentach;
    - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
    - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

    Proponowana liczba osób w grupie: 1.

12. Zbadanie funkcjonalności dostępnej w środowisku R dotyczącej predykcji przestrzennej metodą kriging (uniwersalny, zwykły, prosty), predykcji IDW (IDW - inverse distance weighted interpolation) oraz regresji liniowej. Celem zadania jest ocena dostępnych metod pod kątem ich przydatności do analizy danych przestrzennych: poprawność działania, czas wykonania w zależności od wielkości danych, łatwość użycia, dostępność materiałów pomocniczych.

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;

- zbiory danych i opis danych wykorzystanych w eksperymentach;
- wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
- skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

Proponowana liczba osób w grupie: 2.

13. Zbadanie funkcjonalności dostępnej w środowisku R dotyczącej analiza danych punktowych - funkcje K, L, G. Celem zadania jest ocena dostępnych metod pod kątem ich przydatności do analizy danych przestrzennych: poprawność działania, czas wykonania w zależności od wielkości danych, łatwość użycia, dostępność materiałów pomocniczych.

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;
    - zbiory danych i opis danych wykorzystanych w eksperymentach;
    - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
    - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

    Proponowana liczba osób w grupie: 1.

14. Zbadanie funkcjonalności dostępnej w środowisku R dotyczącej regresji przestrzennej, w tym autoregresji. Celem zadania jest ocena dostępnych metod pod kątem ich przydatności do analizy danych przestrzennych: poprawność działania, czas wykonania w zależności od wielkości danych, łatwość użycia, dostępność materiałów pomocniczych.

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;
    - zbiory danych i opis danych wykorzystanych w eksperymentach;
    - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
    - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

    Proponowana liczba osób w grupie: 1.

15. Zbadanie funkcjonalności dostępnej w środowisku R dotyczącej tworzenia map: w tym map statycznych, animowanych, interaktywnych. Badanie należy przeprowadzić na różnej wielkości danych oraz różnego typu danych (w tym, jeżeli jest to możliwe, na danych dostępnych w środowisku R).

    Dokumentacja projektowa powinna zawierać:
    - opis metod reprezentacji i wizualizacji;
    - opis środowiska, w którym zostały wykonane testy;
    - zbiory danych i opis danych wykorzystanych w eksperymentach;
    - wyniki przeprowadzonych testów, w tym wnioski i ocena (co się udało pokazać daną metodą, gdzie to widać, jaka jest interpretacja);
    - skrypt R, który umożliwi ponowne wykonanie prezentowanych w raporcie badań.

Proponowana liczba osób w grupie: 1.

16. Zadane jest miejsce początkowe i miejsce docelowe podróży. Napisać aplikację, która wskaże miejsca do odwiedzenie (dowolnie wybrane tzw. punkty zainteresowanie -POI) oraz wyznaczy trasę przejazdu przy spełnieniu ograniczeń dotyczącej trasy: wydłużenia czasu przejazdu, długości trasy oraz okresu w czasie podróży, w którym dany punkt ma być odwiedzony (np. po godzinie od rozpoczęcia podróży).

    Aplikacja powinna zawierać autonomiczny moduł do pokazania wyznaczonej trasy na mapie. Do napisania aplikacji należy wykorzystać dostępne serwisy oferujące dane przestrzenne (np. serwis OpenStreetMap http://www.openstreetmap.org/).

    Proponowana liczba osób w grupie: 2
    Dokumentacja końcowa powinna zawierać:
    - Opis metody wyszukiwania.
    - Wyniki eksperymentów pozwalające na ocenę jakości wyników wyszukiwania (precyzję i kompletność).
    - Informacje o implementacji (skomentowany kod źródłowy, zastosowany model danych, architektura aplikacji).

17. Wyszukiwanie miejsc o zadanych cechach - napisać aplikację do wyszukiwania miejsc na podstawie podanych przez użytkownika kryteriów, m.in:
    - odległości w km oraz czasu dojazdu;
    - odległości docelowego miejsca od innych obiektów (np. od jeziora)

    Aplikacja powinna umożliwiać pokazanie znalezionych miejsc na mapie.
    Do napisania aplikacji należy wykorzystać dostępne serwisy oferujące dane przestrzenne (np. serwis OpenStreetMap http://www.openstreetmap.org/).

    Proponowana liczba osób w grupie: 1-2
    Dokumentacja końcowa powinna zawierać:
    - Opis metody wyszukiwania
    - Wyniki eksperymentów pozwalające na ocenę jakości wyników wyszukiwania (precyzję i kompletność).
    - Informacje o implementacji (skomentowany kod źródłowy, zastosowany model danych, architektura aplikacji).

Literatura

Analiza danych przestrzennych w R

- Lovelace, Robin, Jakub Nowosad, and Jannes Muenchow. Geocomputation with R. Chapman and Hall/CRC Press, 2019 (https://geocompr.robinlovelace.net/)
- Bivand, Roger S., et al. *Applied spatial data analysis with R*. New York: Springer, 2013 (książka jest dostępna online w bibliotece PW)
- Brunsdon, Chris, and Lex Comber. *An introduction to R for spatial analysis and mapping*. Sage, 2015.

Grupowanie

- Guha S., Rastogi R., Shim K., *ROCK: A robust clustering algorithm for categorical attributes*, Proceedings of the International Conference on Data Engineering, Sydney 1999, pp. 512–521.
- Karypis G., Han E., Kumar V., *CHAMELEON: A hierarchical clustering algorithm using dynamic modeling*, IEEE Computer: Special Issue on Data Analysis and Mining, vol. 32, no. 8, 1999, pp. 68–75.
- Ng R. T. , Han J., *Efficient and effective clustering methods for spatial data mining*, Proc. 20th Int. Conf. on Very Large Data Bases, Morgan Kaufmann, Santiago 1994, pp. 144–155.
- Ester M., Kriegel H.–P., Sander J.: *Algorithms and applications for spatial data mining*, Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001
- Estivill–Castro V., Lee I.: *AMOEBA: hierarchical clustering based on spatial proximity using Delaunay diagram*, In Proceedings of the 9th International Symposium on Spatial Data Handling, 2000
- Estivill–Castro V., Lee I.: *AUTOCLUST: automatic clustering via boundary extraction for mining massive point–data sets*, Proceedings of the 5th International Conference on Geocomputation, 2000
- Zhang T., Ramakrishnan R., Linvy M., *BIRCH: an efficient data clustering method for very large databases*, Proc. ACM SIGMOD Int. Conf. on Management of Data, ACM Press, 1996, pp. 103–114.

Klasyfikacja

- Koperski K., Han J., Stefanovic N.: *An efficient two–step method for classification of spatial data*, Proc. Int. Symp. on Spatial Data Handling (SDH `98), 1998
- Frank, R., Ester, M. and Knobbe, A. *A multi-relational approach to spatial classification*. Proceedings KDD. 2009. Dostępne także wideo autora (15.02.2011): http://videolectures.net/kdd09_frank_mrasc/.

Asocjacje/Kolokacje

- Xiong H., Shekhar S., Huang Y., Kumar V., Ma X., Yoo J. S., *A framework for discovering co–location patterns in data sets with extended spatial objects*, In Proc. 2004 SIAM International Conference on Data Mining (SDM), 2004.
- Zhang X., Mamoulis N., Cheung D. W., Shou Y., *Fast mining of spatial collocations*, KDD, 2004.
- Koperski K., Han J.: *Discovery of spatial association rules in geographic information databases*, Proceedings of 4th International Symposium on Large Spatial Databases, August 1995
- Morimoto Y.:*Mining Frequent Neighboring Class Sets In Spatial Databases*, KDD'01, San Francisco 2001
- Shekhar S, Huang Y.: *Discovering Spatial Co–Location Patterns: A summary of results*, In Proc of SSTD, Redondo Beach, 2001
- Wan, Y., Zhou, J. and Bian, F. *CODEM: A novel spatial co-location and de-location patterns mining.* Shandong : International School of Software, Wuhan University, 2008, s. 576 - 580.

Wykorzystanie danych z sieci społecznościowych LBSN

- Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, and Elena Baralis. *Predicting your next stop-over from location-based social network data with recurrent neural networks*. In RECSYS 2017, 2nd ACM International Workshop on Recommenders in Tourism (RecTour'17), CEUR Proceedings Vol. 1906, August 27-31, 2017, Como, Italy, Como, ITALIE, 08 2017.
- Lian, Jianxun, et al. "Restaurant survival analysis with heterogeneous information." *Proceedings of the 26th International Conference on World Wide Web Companion*. 2017.