Eksploracja danych
Tematy projektów

1. Zastosowanie algorytmów *Gradient Boosted Decision Trees / Light Gradient Boosting* w klasyfikacji (biblioteka XGBoost)
   - Przykładowe zbiory danych:
     - The Sloan Digital Sky Survey `http://www.sdss.org/` (klasyfikacja obiektów astronomicznych)
     - `http://www.ttss.krakow.pl/internetservice/` (predykcja opóźnienia tramwaju w Krakowie)
     - Spooky Author Identification `https://www.kaggle.com/c/spooky-author-identification/data` (identyfikacja autora)
   - Dopasowywanie parametrów modelu
2. Zastosowanie algorytmów *Gradient Boosted Decision Trees / Light Gradient Boosting* do prognozowania szeregów czasowych (biblioteka XGBoost)
   - Przykładowe zbiory danych:
     - Przewidywanie kursu wybranych kryptowalut
     - Store Sales Forecasting `https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data`
   - Dopasowywanie parametrów modelu
   - Porównanie z modelami: Arima, ESN, Prophet
3. Wykrywanie i wizualizacja społeczności w grafach Twittera (followers, mentions)
   - Przygotowanie zbioru danych:
     - Stanford Large Graph Database Collection

    - Próbkowanie grafu Twittera za pomocą API (co najmniej 2M wierzchołków)
  - Algorytmy
    - Label propagation
    - Infomap
    - Walktrap
    - Spinglass
  - Opis społeczności za pomocą najczęściej używanych hashtagów
4. Wykrywanie newralgicznych fragmentów sieci drogowej w oparciu o miary centralności grafu
  - Zbiory danych:
    - Sieć drogowa Kaliforni `https://snap.stanford.edu/data/roadNet-CA.html`
    - Sieć drogowa Krakowa lub wybranego regionu Polski (OpenStreetMap database)
  - Miary centralności:
    - Betweenness centrality
    - Random-walks betweenness centrality
    - PageRank
    - Local Clustering Coefficient
  - Identyfikacja potencjalnych nowych krawędzi, które ustawnią ruch
5. Wykrywanie społeczności w grafie współautorstwa DBLP
  - Zbiór danych:
    - `https://dblp.uni-trier.de/faq/How+can+I+download+the+whole+dblp+dataset`
    - `https://snap.stanford.edu/data/com-DBLP.html`
  - Weryfikacja skuteczności działania wybranych algorytmów wykrywania społeczności w grafie w oparciu tytuły journala
  - Przedstawienie deskryptorów topologicznych grafu współautorstwa
  - Algorytmy
    - Label propagation
    - Infomap

  - Walktrap
  - Spinglass

6. Analiza danych z Krajowego Rejestru Sądowego
   - `https://rejestr.io/api-krs`
   - Na podstawie listy organizacji skonstruuj grafy odzwierciedlające:
     - Organizacje powiązane co najmniej jednym członkiem reprezentacji
     - Osoby powiązane organizacją, którą wspólnie reprezentują
   - Przedstaw analizę ilościową grafów j.w. wyznaczając:
     - Największe społeczności w grafie
     - Największe kliki w grafie
     - Rozkład liczby wierzchołków w spójnych składowych grafu
     - Rozkład i największe wartości *clustering coefficient*
   - Przedstaw interaktywną wizualizację największych spójnych składowych grafu
7. Analiza danych radiacynych Safecast
   - Zbiór danych:
     - `https://github.com/Safecast/safecastapi/wiki/Data-Sets`
   - Zbadaj korelację pomiędzy wysokością nad poziom morza a poziomem radiacji
   - Przedstaw ilościową analizę i wizualizację zmian radiacji w Prefekturze Fukushima w Japonii (od początku dostępu danych pomiarowych)
   - W oparciu o dane historyczne zbuduj model regresji pozwalający przewidzieć poziom radiacji w następnych latach (w danym obszarze)
8. Analiza danych o zanieczyszczeniu powietrza (Airly) oraz danych pogodowych (Ecowitt)
   - Zbiór danych:
     - `https://developer.airly.eu/docs#endpoints.measurements`
   - Wykorzystaj dane z czujników jakości powietrza do segmentacji obszaru Krakowa na podobszary charakteryzujące się podobną dynamiką zmian jakości powietrza

- Na podstawie danych historycznych oblicz następujące wskaźniki ilościowe jakości powietrza (na podobszar):
  - Najdłuższy ciągły okres, w którym norma została przekroczona w danych podobszarze (rozdzielczość godzinowa)
  - Najgorsza średnia jakość powietrza w danym okresie czasu
- Zaproponuj metodę pozwalającą na modyfikację rozdzielczości podobszarów (klastrów)
- Przedstaw wizualizację wyników analizy

9. Analiza danych o użyciu rowerów / miejskich skuterów elektrycznych
   - Zbiór danych:
     - `https://data.world/cityofchicago/2kfw-zvte`
     - `https://data.world/louisville/83299427-e232-49a1-8cfb-d524260d6673`
     - `https://data.world/minneapolismn/4001ebcfd38d404c9662cbfc3311c3c8-0`
     - `https://data.edmonton.ca/Transportation/E-Scooter-Share-API/q9ny-crw9/data?pane=feed`
     - `http://dev.cityofchicago.org/open%20data/2019/07/17/scooter-gbfs-public-feeds.html`
   - Wykorzystaj dane o lokalizacji, stanie baterii, przemieszczeniu i ostatniej aktywności do przedstawienia zaawnsowanych statytyk ruchu jednośladów:
     - Trendy dobowe
     - Rozkład wielkości przemieszczeń
     - Najczęściej, najrzadziej używane pojazdy
     - Obszary największej i najmniejszej aktywności transportowej w mieście
     - Obserwacje odstające (outliers)
   - Przedstaw wizualizację wybranych wyników analizy na mapie
10. Badanie genetycznych i metabolicznych powiązań między organizmami
    - Zbiór danych:
      - `https://metacyc.org`
      - `http://bigg.ucsd.edu/models`
      - `http://konect.uni-koblenz.de`

- Wykorzystaj dane o ścieżkach metabolicznych do budowy grafów sieci metabolicznych dla zbioru mikroorganizmów z bazy danych MetaCyc
- Zapropnuj metodę konstrukcji drzewa filogenetycznego wykorzystującą podobieństwo strukturalne sieci metabolicznych i przedstaw wizualizację wyników

11. Analiza podobieństw między państwami na podstawie danych ekonomicznych i socjologicznych Banku Światowego

    - Zbiór danych:
        - World Bank Open Data
    - Wykorzystaj otwarte dane historyczne z Banku Światowego do zbudowania funkcji podobieństwa między państwami wykorzystującej różne parametery rozwoju (ostatnie 30 lat)
    - Przedstaw wizualizację wyników (t-SNE)
    - Przedstaw przykładowe klasteryzacje zbioru danych

12. Algorytmy klasyfikacji grafów w analizie sieci społecznościowych oraz chemii molekularnej

    - Niezmienniki grafowe
    - `https://chrsmrrs.github.io/datasets/docs/datasets/`
    - `https://ogb.stanford.edu/docs/graphprop/`
    - `https://arxiv.org/abs/1811.03508`