POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Projekt

DR INŻ. AGATA MIGALSKA

01

TEMATYKA PROJEKTÓW

02

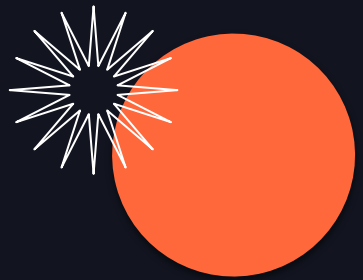
KAMIENIE MILOWE

03

OŚ CZASU

04

NARZĘDZIA PROGRAMISTYCZNE



01

TEMATYKA PROJEKTÓW

TEMATYKA PROJEKTÓW

- Projekt powinien odkrywać ciekawe relacje w ramach znacznej ilości danych
- Rozszerz/popraw/przyspiesz jakiś istniejący algorytmy lub zdefiniuj nowy problem i rozwiąż go.

Grupa projektowa może składać się z 2±1 osób.

PRZYKŁADOWE ŹRÓDŁA DANYCH

- Wikipedia
- Twitter
- Blogi i wiadomości
- Opinie o restauracjach
- Spotify
- Eurostat
- Github
- Dane z urządzeń IoT, zegarków treningowych
- Dane klimatyczne np. Climate Data Online
- ... (wiele, wiele innych)

- Kaggle: https://www.kaggle.com/datasets
- Open ML: https://www.openml.org/

WIKIPEDIA

- W jaki sposób statystyki odwiedzin strony Wikipedii korelują ze zdarzeniami zewnętrznymi, klęskami żywiołowymi? Porównaj występowanie fraz np. na Twitterze i odwiedzin na stronach Wikipedii
- Na podstawie historii edycji stron i dyskusji na Wikipedii
    - Jak ewoluują artykuły? (Możesz np. użyć odległości Levenshteina, aby zmierzyć różnice między wersjami artykułu)
    - Którzy użytkownicy dokonują jakich typów zmian? Jacy użytkownicy dokonują jakich typów zmian?
    - Zaproponuj, który użytkownik powinien edytować stronę
    - Czy edytujący zawsze dyskutują? Czy dyskutujący zawsze edytują? Jak się pokrywają te dwa grafy społeczne
    - Zaproponuj użytkownikom strony do edycji

TWITTER

- Popularne tematy: wzrosty i spadki popularności
- Jaki jest cykl życia hashtagów?
- Znajdowanie influencerów
- Grupowanie tweetów według tematu lub kategorii
- Analiza sentymentu – czy ludzie są nastawieni pozytywnie czy negatywnie?

SPOTIFY

- Jak zmieniał się Twój gust muzyczny w czasie
- Relacje pomiędzy utworami
- Lepszy algorytm rekomendujący utwory np. wybór następnej piosenki na podstawie pory dnia i dotychczasowych preferencji
- ...

02

KAMIENIE MILOWE

ETAPY PROJEKTU ANALITYCZNEGO

KAMIENIE MILOWE PROJEKTU

1. Propozycja projektu (150-250 słów streszczenia projektu)
2. Eksploracyjna analiza danych zawierająca:
    a. informacje o pochodzeniu danych,
    b. opis zmiennych (nazwa, znaczenie, typ),
    c. badanie brakujących i błędnych wartości wraz z rozwiązaniem tych problemów,
    d. analizę rozkładów zmiennych,
    e. analizę zależności między zmiennymi,
    f. inne interesujące obserwacje na podstawie danych.
3. Zastosowanie jednej lub wielu technik eksploracji danych (klasyfikacja / regresja / klastrowanie / uczenie asocjacji) do rozwiązania postawionego problemu.

03

OŚ CZASU

OŚ CZASU

- 13/14.10 - omówienie pomysłów na zajęciach / rozwiązywanie problemów z instalacją oprogramowania
- 20/21.10 - Milestone 1 Deadline: Przesłanie Propozycji projektu (PDF) zawierającej:
  - tytuł,
  - skład grupy projektowej i ew. nazwa grupy,
  - opis problemu, którego projekt ma dotyczyć wraz ze wskazaniem źródeł danych, które zostaną wykorzystane (150-250 słów).
- 17.11/25.11 - Milestone 2 Deadline: Przedstawienie otrzymanych wyników (w formie konsultacji na zajęciach) oraz przesłanie raportu (PDF) podsumowującego ten etap.
- 15.12/16.12 - Milestone 3 Review: Przedstawienie otrzymanych wyników (w formie konsultacji na zajęciach) oraz przesłanie raportu (PDF) podsumowującego dotychczasowe wyniki etapu.
- 19/20.01 - Milestone 3 Deadline: Złożenie pełnego raportu z projektu (PDF) (zawierającego wszystkie części z uwzględnieniem uwag zgłoszonych do poprzednich części).
- 19/20.01 oraz 26/27.01 - Prezentacje multimedialne z projektów przed resztą grupy.

KRYTERIA OCENY

- Raporty - 70pkt:
  - Propozycja projektu - 10pkt
  - Analiza eksploracyjna danych - 10pkt
  - Częściowy raport z eksploracji danych - 10pkt
  - Raport końcowy - 40pkt
- Prezentacja - 30pkt
- Dodatkowo 2 nagrody publiczności (+10 pkt każda):
  - za najlepszą prezentację
  - za najciekawszy pomysł

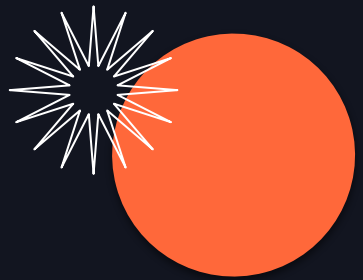
Każdy dzień (kalendarzowy) zwłoki w złożeniu raportu kosztuje 1/10 możliwych do otrzymania punktów.
Za koniec dnia uznaje się północ wg czasu obowiązującego w Polsce w danym dniu.
Arbitrem jest zegar na serwerze pocztowym.

PREZENTACJA KOŃCOWA

Zawartość prezentacji końcowej:

1. Cel projektu - postawienie pytania lub problemu do rozwiązania
2. Dane - źródła, zawartość, analiza kompletności, sposób poradzenia sobie z brakującymi danymi, czyszczenie danych, tworzenie cech
3. Eksploracja danych (wnioski, obserwacje)
4. Rozwiązanie postawionego problemu - predykcja, klastrowanie, metody ekstrakcji.
5. Otrzymane wyniki
6. Wnioski

Czas trwania prezentacji: 20min
Q&A: 10min



04

NARZĘDZIA PROGRAMISTYCZNE

NARZĘDZIA PROGRAMISTYCZNE

1. Języki programowania: Python, ew. R lub Julia
2. Biblioteki (Python):
    a. NumPy, SciPy, Pandas (Scientific Python Ecosystem)
    b. Scikit-learn (algorytmy uczenia maszynowego)
    c. Matplotlib / Plotly / Seaborn (wizualizacje)
    d. jupyterlab (instaluje wszystkie zależności zarówno dla Jupyter Notebook jak i Jupyter Lab)
3. Środowisko deweloperskie: Jupyter Notebook, IDE (visual studio code, PyCharm, DataSpell)
4. Raporty: Jupyter Notebook (eksport do PDF), dowolny biurowy edytor tekstu, dowolny edytor Markdown (można ładnie wyrenderować taki raport na githubie)

KONTAKT

agata.migalska@pwr.edu.pl

THANKS!

DZIĘKUJĘ ZA UWAGĘ