

Wybrane zagadnienia sztucznej inteligencji Analiza dużych zbiorów danych metodą k-średnich

Przebieg ćwiczeń laboratoryjnych i warunki zaliczenia przedmiotu
semestr VI rok akademicki 2023/24

Ćwiczenia laboratoryjne

- na studiach dziennych 8 zajęć * 3 godziny = 24 godziny.

Przedmiotem ćwiczeń laboratoryjnych jest analiza dużych zbiorów danych **metodą k-średnich**.

Studenci otrzymują trzy moduły, autorstwa prowadzącego przedmiot, zawierające definicje funkcji niezbędnych do realizacji algorytmu metodą **k-średnich** zapisane w języku **Python 3.9**. Oprócz tego: duży zbiór danych **TopBabyNamesbyState** w postaci 10 500 pięcio-atrybutowych rekordów zapisanych w formacie **.csv**.

Studenci pracują w jedno- lub dwuosobowych zespołach.

Ćwiczenia laboratoryjne przebiegają według niżej podanego schematu:

1. Zajęcia nr 1 i 2:

Omówienie projektu analizy dużych zbiorów danych metodą k-średnich w oparciu o wykład wstępny i plik omawiający przebieg zajęć i zaliczeń.

Praca w domu między zajęciami nr 2 i 3. W tym czasie studenci kontaktują się z prowadzącym przedmiot poprzez Teams. Zapoznanie się szczegółowo z ideą i działaniem metody k-średnich w oparciu o wykład wstępny. Zapoznanie się ze metodą normalizacji danych przed rozpoczęciem ich analizy, oraz budową i własnościami dostarczonego programu w języku **Python 3.9**.

2. Zajęcia nr 3 i 4 :

Badania własności dostarczonej przykładowej bazy danych za pomocą dostarczonego oprogramowania. Będzie to przede wszystkim **analiza wpływu wybranej ilości klastrów na możliwości wydobywania wiedzy z poszczególnych klastrów**. Odbywać się to będzie poprzez pokazanie w wersji elektronicznej i dyskusja z prowadzącym zajęcia otrzymanych wyników. Wymagane będzie:

- **dobre rozumienie działania algorytmu k-średnich, oraz metody normalizacji danych,**
- **rozumienie własności, struktury i wzajemnych zależności między atrybutami dostarczonego zbioru danych.**

Na zajęciach tych może mieć miejsce **zaliczenie etapu pierwszego**. Na zaliczeniu etapu I student może otrzymać **10-20 punktów** (jeśli zaliczy I-szy etap), lub 0 punktów w przypadku nie zaliczenia. Ocena będzie indywidualna, mimo ewentualnej pracy w dwuosobowym zespole.

3. Zajęcia nr 5 i 6:

Przed zajęciami nr 5 studenci mają **obowiązek pozyskania z Internetu** do dalszych badań innej, odpowiednio dużej bazy danych opisującej realny obszar rzeczywistości.

Do zadań studentów na ćwiczeniach nr 5 i 6 należeć będzie (w kontakcie z prowadzącym przedmiot):

- **przeprowadzenie przed rozpoczęciem zajęć nr 5 i 6 normalizacji danych** (normalizacja wymaga zatwierdzenia przez prowadzącego zajęcia),
- **przeprowadzenie badań wpływu wyboru liczby klastrow na efektywność algorytmu z wyborem optymalnej liczby klastrow,**
- **przeprowadzenie badań i analiza wpływu metody liczenia odległości (euklidesowa, Manhattan) na efektywność algorytmu.**
- **umiejętność prezentacji i interpretacji osiągniętych wyników.**

4. Zajęcia nr 7 - 8:

Zajęcia przeznaczone są na ocenę i dyskusję osiągniętych przez poszczególne osoby wyników, oraz końcową ocenę indywidualną studentów (etap drugi).

Do zaliczenia należy przygotować **sprawozdanie końcowe** w wersji papierowej.

Za wykonania etapu drugiego student może otrzymać **20-40 punktów**.

Promowana będzie: jakość i terminowość realizacji poszczególnych etapów projektu, organizacja i współdziałanie poszczególnych członków w zespole, wartość osiągniętych rezultatów, przy czym brany będzie pod uwagę indywidualny wkład każdego z członków zespołu. Wcześniejsze zaliczenie etapu drugiego promowane będzie dodatkowymi punktami.

Przedmiotem oceny etapu drugiego będzie jednak przede wszystkim: **umiejętność dostosowania dostarczonego oprogramowania do własności nowej bazy danych**, ale też - sposób prezentacji osiągniętych wyników w sprawozdaniu końcowym, dokumentującym osiągnięte wyniki.

Zawartość sprawozdania z laboratorium "Sztucznej inteligencji" na temat : **"Inteligentna analiza dużych zbiorów danych"**:

- strona tytułowa,
- zbiór danych **etapu drugiego**,
- moduły kodów programu w języku *Python* dostosowane do aktualnego zbioru danych,
- wyniki badań **wpływu liczby klastrow**, oraz **wyboru metody liczenia odległości** między rekordami danych (**odległość euklidesowa, odległość Manhattan**) na **efektywność wydobywania wiedzy ze zbioru danych**.

Warunki zaliczenia laboratorium

Za ćwiczenia laboratoryjne student może uzyskać na zaliczeniu maksymalnie 60 punktów. Ćwiczenia uznaje się za zaliczone, jeśli student uzyska **nie mniej niż 30 punktów**.

Tabela ocen końcowych:

30 – 35 pk. ocena 3.0

36 – 41 pk. ocena 3.5

42 – 47 pk. ocena 4.0

48 – 53 pk. ocena 4.5

54 – 60 pk. ocena 5.0

Wyróżniające się prace będą oceniane na 6.0

Ostatni dzień zajęć laboratoryjnych, wskazany w harmonogramie zajęć, jest ostatecznym terminem zaliczenia ćwiczeń.