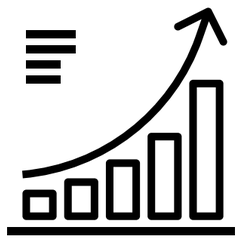
магістерська дисертація на тему

ІНТЕЛЕКТУАЛЬНА СИСТЕМА КЛАСТЕРИЗАЦІЇ ПОМИЛОК ЯК СКЛАДОВА АВТОМАТИЗАЦІЇ ТЕСТУВАННЯ

Автор: студентка
групи КА-з82мп
Глушко С. М.

АКТУАЛЬНІСТЬ



ріст
сфери IT
в Україні

важливість
забезпечення
якості

велика
кількість
тестів

значний час
аналізу
результатів

більшість засобів
для агрегації
результатів тестування:

не забезпечують
інтелектуального
аналізу помилок

ReportPortal:

заявлено:
класифікація

за відгуками:
працює некоректно

ПОСТАНОВКА ЗАДАЧІ

Створити систему, яка здатна
кластеризувати отриманий набір
результатів виконання тестів,
поданий у загальновживаному форматі,
на основі помилок. а саме
повідомлень про помилки.

МЕТА

розв'язок часткового випадку задачі кластеризації даних, а саме кластеризації помилок тестування

ОБ'ЄКТ

інтелектуальний аналіз результатів тестування

ПРЕДМЕТ

системи кластеризації помилок тестування

ВЕКТОРИЗАЦІЯ ТЕКСТУ

BAG OF WORDS	TF IDF	WORD 2 VEC

Висновок: найкраще для розв'язку обраної задачі підходить алгоритм TF-IDF

АЛГОРИТМИ КЛАСТЕРИЗАЦІЇ

Нейронний:
карти Кохонена

Центроїдний:
к-середніх

Статистичний:
EM-алгоритм

Ієрархічний:
агломеративний

Щільнісний:
DBCSAN

Графовий:
HCS

Необхідність апріорних знань

Здатність коректно працювати з викидами

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЗА КРИТЕРІЯМИ

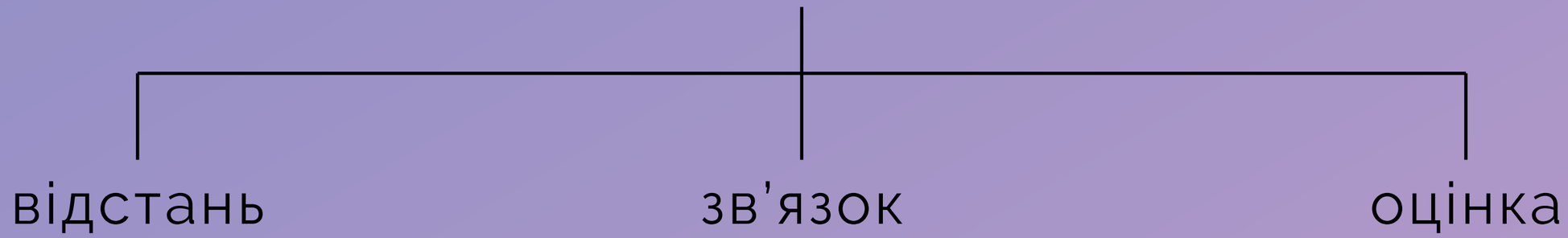
Залежність від ініціалізації

Обчислювальна складність

АГЛОМЕРАТИВНИЙ АЛГОРИТМ

послідовно об'єднує менші кластери в більші
до отримання одного великого кластеру,
що включає всі об'єкти

МОДИФІКАЦІЇ



ВІДСТАНЬ:

- евклідова:

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- мангеттенська:

$$d(X,Y) = \sum_i |x_i - y_i|$$

ЗВ'ЯЗОК:

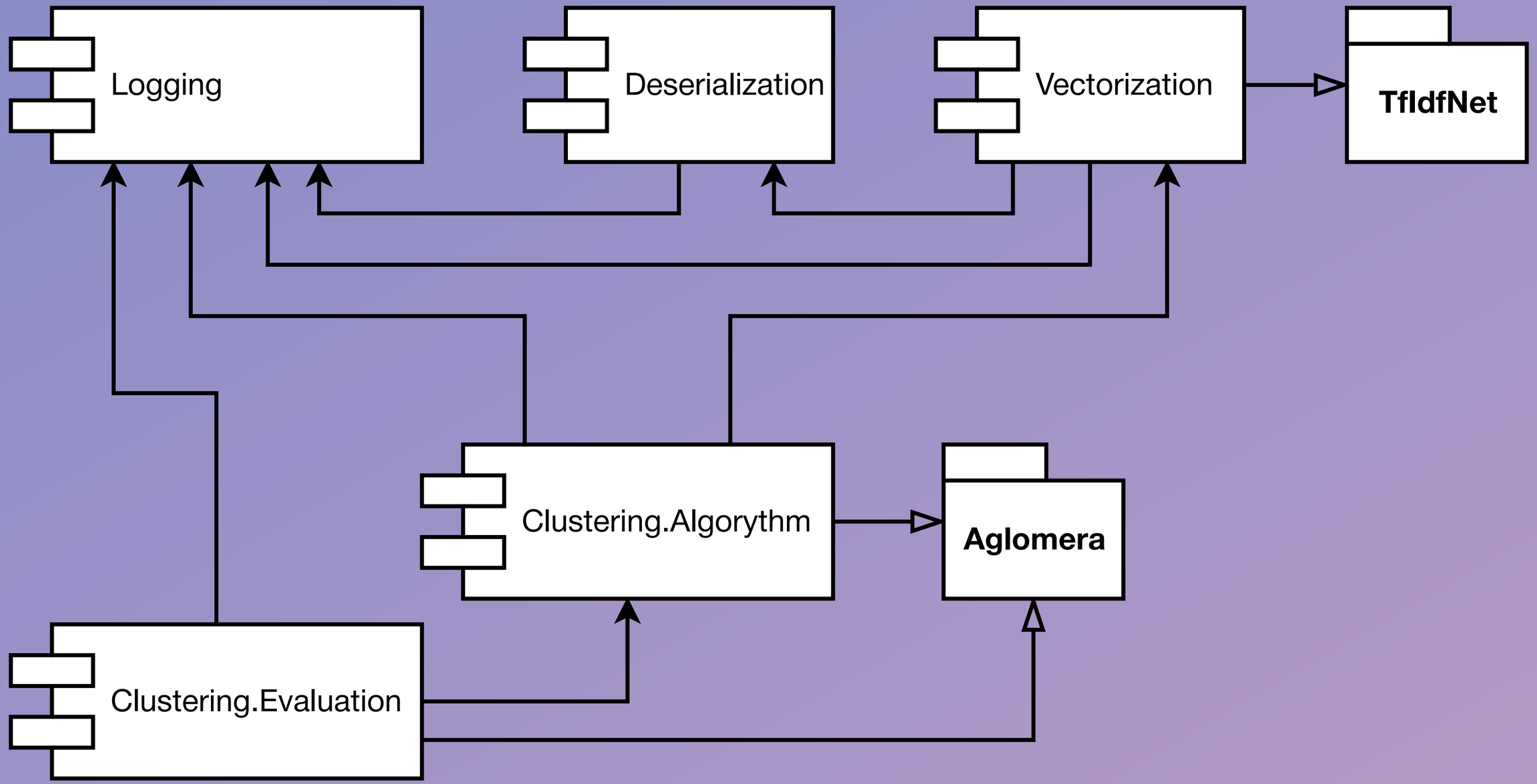
- середній: $dist(A,B) = average(dist(a,b), a \in A, b \in B)$
- повний: $dist(A,B) = max(dist(a,b), a \in A, b \in B)$
- одинарний: $dist(A,B) = min(dist(a,b), a \in A, b \in B)$

ОЦІНКА:

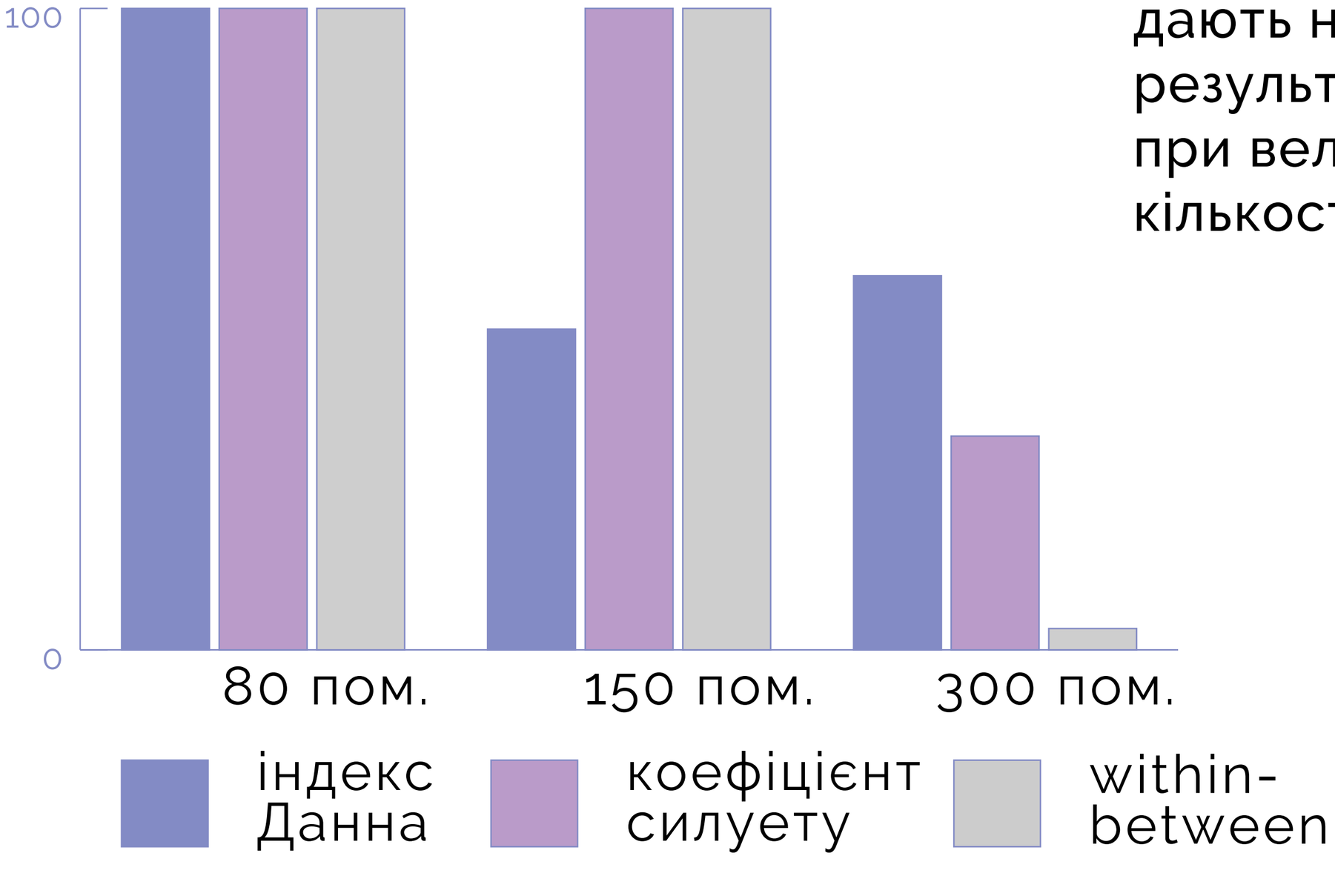
- коефіцієнт силуету
- індекс Данна
- within-between

ВИМОГИ ДО ФУНКЦІОНАЛУ

- серіалізувати вхідні дані для отримання об'єктів
- векторизувати текстові повідомлення для отримання числових векторів для кластеризації
- отримати розбиття на кластери для знаходження закономірностей в даних
- оцінити розбиття на кластери для знаходження найкращого
- вести логування, щоб знати прогрес роботи програми


Logging
Deserialization
Vectorization
TfldfNet
Clustering.Algorythm
Aglomera
Clustering.Evaluation

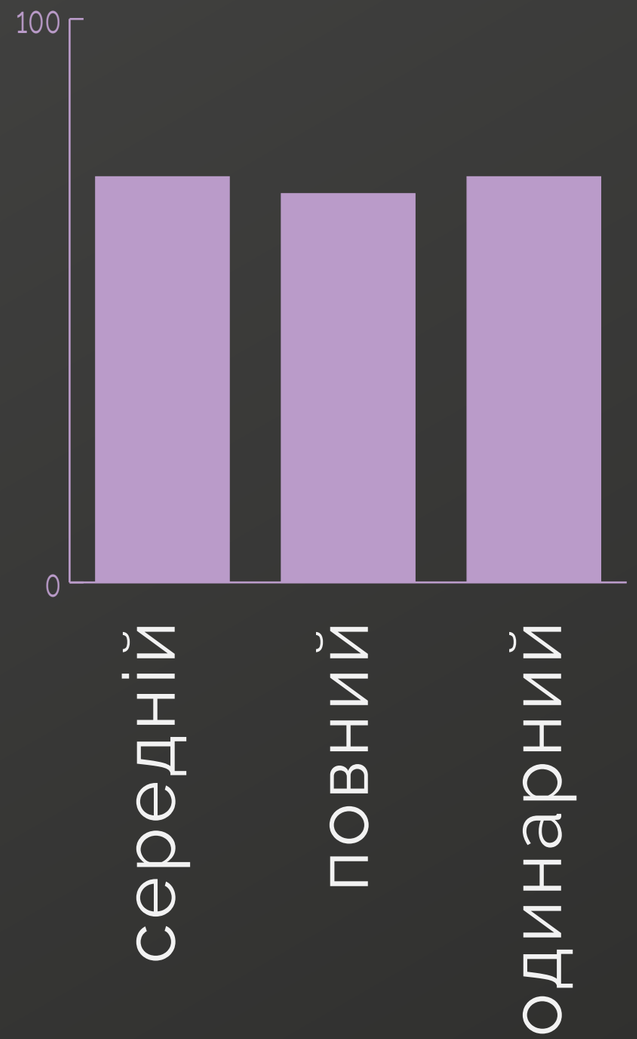
Середня евристична оцінка обраного за критерієм якості розбиття



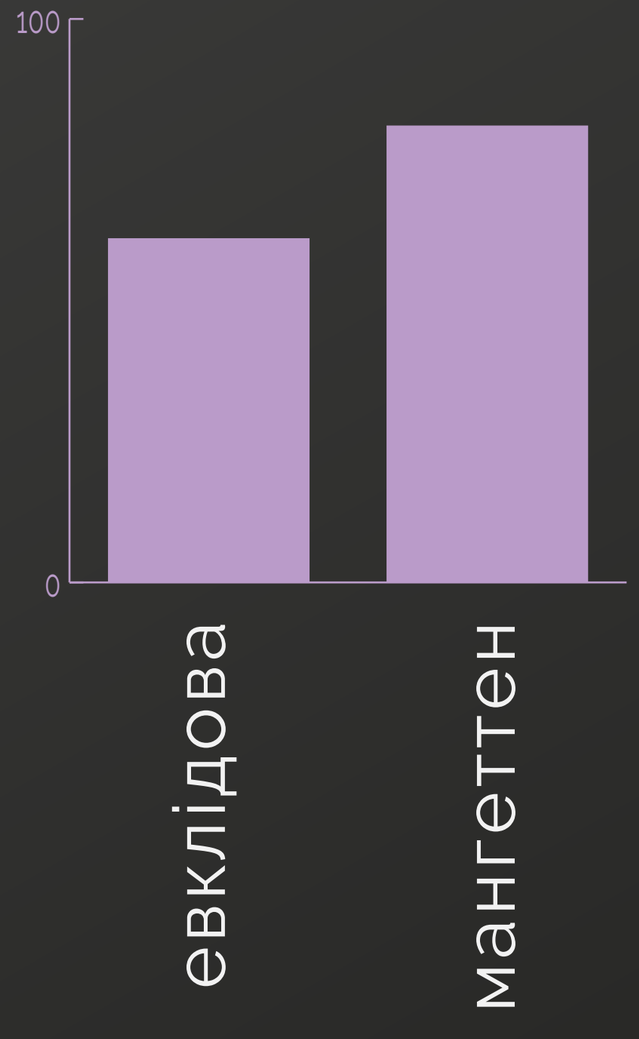
Висновок: **13**

усі критерії
дають незадовільний
результат
при великій
кількості помилок

Якість результату залежно від типу зв'язку - практично не залежить



Якість результату залежно від метрики відстані - мангеттенська підходить краще



ПРАКТИЧНА ЦІННІСТЬ ТА ПОДАЛЬШИЙ РОЗВИТОК

розроблена система кластеризації помилок дозволила зменшити час аналізу результатів автоматизованого тестування на

30%

- підбір іншого критерію якості розбиття для роботи із більшою кількість помилок
- реалізація API сервісу
- підтримка вхідних даних у форматі JSON

ЗАПИТАННЯ ТА ВІДПОВІДІ