**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА**

Факультет прикладної математики та інформатики

(повне найменування назва факультету)

Кафедра інформаційних систем

(повна назва кафедри)

**Магістерська робота**

Розробка системи кластеризації антитіл на основі коефіцієнту перехресного зв’язування

Виконав: студент групи ПМІМ-22с

спеціальності

122-“Комп’ютерні науки”

(шифр і назва спеціальності)

Зелінський О. А.

(підпис) (прізвище та ініціали)

Керівник Горлач В. М.

(підпис) (прізвище та ініціали)

Рецензент

(підпис) (прізвище та ініціали)

Львів – 2022

# ЗМІСТ

[ЗМІСТ 2](#_Toc118840202)

[Вступ 3](#_Toc118840203)

[1 Теоретичні відомості 4](#_Toc118840204)

[2 Опис даних та алгоритмів 7](#_Toc118840205)

[2.1 Опис даних 7](#_Toc118840206)

[2.2 Алгоритм роботи з категоріальними змінними 9](#_Toc118840207)

[2.3 Алгоритм роботи з кількісними змінними 12](#_Toc118840208)

[3 Практична реалізація 15](#_Toc118840209)

[4 Результати 21](#_Toc118840210)

[4.1 Результати роботи алгоритм для категоріальних змінних 23](#_Toc118840211)

[4.1 Результати роботи алгоритм для кількісних змінних 25](#_Toc118840212)

[Висновок 32](#_Toc118840213)

[Список Використонаї Літератури 33](#_Toc118840214)

# Вступ

Епідемія Covid-19 показала, що людству досі досить важко контролювати гострі респіраторні вірусні інфекції та боротися з ними. За даними ВООЗ, майже 613 мільйонів людей у всьому світі були інфіковані Covid-19 і понад 6,5 мільйонів людей померли внаслідок цієї хвороби.

Однак загальновідомо, що це була не перша і, мабуть, не остання така пандемія. Тому вкрай важливо проводити дослідження якомога швидше, щоб захворювання можна було легко виявити та вилікувати. Наступний крок – розробка вакцин, а також тестів, які показують кількість антитіл до того чи іншого вірусу. Зрозуміло, що швидке виявлення захворювання допомагає ізолювати розповсюдження вірусу та ефективніше лікувати хворого, а вакцинація підвищує імунітет до конкретного вірусу та знижує ймовірність негативних (зокрема летальних) наслідків.

У наш час комп’ютери є дуже потужним інструментом, який дозволяє вирішувати не тільки математичні задачі, а й біологічні, хімічні та медичні. Для цього використовуються різні типи моделей і алгоритмів (включаючи алгоритми машинного навчання). Крім того, використання комп’ютерів допомагає вченим скоротити кількість експериментів і рутинної роботи в лабораторіях по всьому світу.

Мета цієї роботи – розглянути проблему пошуку двох оптимальних антитіл до будь-якого вірусу (наприклад, SARS-CoV-2) і запропонувати можливі шляхи її вирішення за допомогою алгоритмів машинного навчання, а точніше алгоритмів кластеризації.

# 1 Теоретичні відомості

Вірус(від. лат. virus – отрута) – це не клітинні форми життя, які є внутрішньо клітинними абсолютними паразитами. Оскільки віруси вражають всі клітинні організми та спричиняють масові захворювання (епідемії) [ книжка з біології 9 клас], людство розробило заходи боротьби з цими захворюваннями серед яких:

1. Ізоляція хворих організмів від здорових (карантин);
2. Лікування за допомогою хімічних антивірусних препаратів (хіміотерапія);
3. Профілактичні щеплення для підвищення стійкості організму (імунізація).

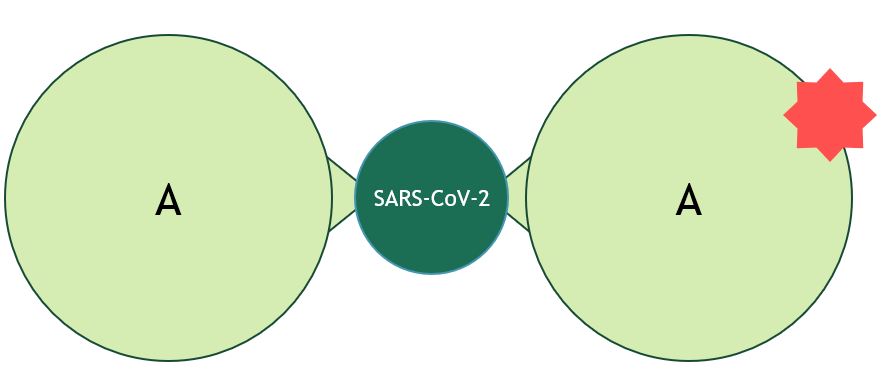
В свою чергу людський організм має власну захисну систему – імунітет, складовими якого є лімфоцити, антитіла, тощо.

Антитіла (імуноглобуліни (Ig)) – це речовини, що утворюються в організмі людей внаслідок введення в нього сторонніх білків, бактерій, отрут та інших речовин, з метою знищити або нейтралізувати потенційну небезпеку. [енциклопедія][словник]

У задачі, що буде розглянута у цій магістерській роботі, дано молекулу протеїну (нуклеопротеїну), яка є загально визнаною ціллю для діагностики SARS-CoV-2 (не для вакцини), що використовується практично всіма швидкими тестами на антиген. До цієї молекули приєднуються два антитіла (вони можуть бути як різними так і однаковими), для того щоб можна було їх відрізнити одне з них помічається \*. Ці два антитіла формують „сендвіч”, це стандартна схема визначення будь-якої білкової речовини.

Для простоти вважатимемо, що все відбувається на площині, антитіла це два круги однакового розміру, що мають невеликий „дзьоб” для взаємодії з вірусом. Антитіла в свою чергу приєднуються до меншого круга який представляє молекулу протеїну. Схематичне зображення цього процесу можна побачити на рисунку 1.1.

У нашому випадку конкуренція йде між антитілами з молярною вагою180 кДа за зв’язування з вірусом з молярною вагою 45 кДа. З цієї причини відбувається запекла конкуренція.

  
Рисунок 1.1 – Модель приєднання антитіл до вірусної молекули

Основне завдання полягає у знаходженні двох антитіл (можуть бути однаковими), таких що знаходяться на оптимальній відстані одне від одного тобто не перетинаються та не знаходяться занадто близько один до одного, щоб почати конкурувати між собою.

Зручним теоретичним методом вирішення цієї проблеми є розбиття списку антитіл на групи за ознакою того наскільки вони заважають один одному або іншими словами чи приєднуються вони до вірусу в одній і тій же області. Те що антитіла належать до різних груп означає, що вони прив’язуються в різних областях та взаємодіють краще ніж якби вони були з однієї.

В сучасних комп’ютерних науках популярним способом для розбиття даних на групи є група алгоритмів, яка називається кластерним аналізом.

Кластерний аналіз (англ. Data clustering) – задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Метою кластерного аналізу є класифікація об’єктів на відносно гомогенні (однорідні) групи, виходячи із досліджуваної кількості ознак (показників, змінних). Об’єкти в групі є відносно подібними з огляду на досліджувані показники і відрізняються від об’єктів у інших групах [методичка з екології]. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя.

Поняття кластеру тяжко однозначно визначити, тому існує доволі велика кількість алгоритмів кластеризації. Оскільки деякі алгоритми працюють зі схожими, або навіть однаковими об’єктами, які вони називають кластерами, то такі об’єкти було названо „кластерні моделі”, та поділено на групи, такі як:

1. Моделі зв’язності
2. Центроїдні моделі
3. Статистичні моделі
4. Моделі засновані на щільності
5. Групові моделі
6. Графові моделі
7. Нейронні моделі

В даній роботі буде розглянуто перші дві моделі, а саме модель зв’язності для алгоритмів ієрархічної кластеризації та центроїдну модель для алгоритму k-means та його модифікації k-mods.

# 2 Опис даних та алгоритмів

## 2.1 Опис даних

Дані для виконання роботи були надані компанією Xema OY, Лаппеенранта, Фінляндія. Для їх отримання був проведений експеримент який складався з декількох кроків, серед яких:

1. Лабораторні дослідження (дослідження прямого інгібування зв’язування)
   1. **Кон’югація HRP з моноклональними антитілами (мАт)**. Для цього використовували один з найпопулярніших методів кон’югації HRP (Faizyme, SAR) з антитілами, вперше описаний Nakane і Kawaoi в 1974 році [стаття Накане]. Періодично окислений HRP утворює ковалентний зв’язок з мАТ після відновлення основи Шиффа бор гідридом натрію.
   2. **Пряме зв’язування мАТ з варіантами N-Ag**. Препарати N-Ag розводили до 0,1 мкг/мл карбонатним буфером рН 9,5. Сто мікролітрів розчину вносили в лунки полістиролового мікропланшета з високою адсорбційною здатністю (KHB, Китай) та інкубували протягом ночі при температурі +4С. Після видалення вмісту мікролунок вакуумом, мікролунки одноразово відмивали промивним розчином для ІФА - 0,1% Tween 20 (Serva, Німеччина) в 0,9% хлориді натрію (Merck, Німеччина) і заливали блокуючим розчином для ІФА (0,1М фосфатний буфер, що містить 0,9% NaCl і 0,5% гідролізованого казеїну) на 2 години при кімнатній температурі, а потім висушували при кімнатній температурі протягом 48 годин. МАТ розводили буфером для ІФА (0,1М фосфатний буфер, що містить 0,9% NaCl і 0,1% гідролізованого казеїну) до рівномірної концентрації 1 мкг/мл. Сто мл розчину mAb інкубували в лунках протягом 30 хвилин при 37С. Лунки тричі промивали промивним розчином для ІФА і додавали HRP-кон'югований овечий антимишачий Ig-HRP кон'югат (Cat# AS302-HRP, Xema) в робочому розведенні в лунки ще на 30 хвилин при 37С. Після 5 відмивань промивним розчином для ІФА в лунки вносили хромогенний субстрат TMB (Cat#R055, Xema) на 15 хвилин, реакцію зупиняли додаванням 5% сірчаної кислоти і вимірювали оптичну щільність при 450 нм (OD450) на ридері для мікропланшетів HiPo (Biosan, Латвія).
2. Перехресне інгібування mAb шляхом прямого зв’язування з твердофазним N-Ag.

N-Ag повної довжини наносили на поверхню полістирольних лунок у концентрації 0,5 мкг/мл (див. попередній параграф). У попередньому тесті кожен HRP-кон’югований мАт серійно розводили (у 10 разів) у мікролунках від 1:100 до 1:1 млн. та інкубували протягом 30 хвилин при 37С. Потім реакцію завершували відмиванням, додаванням субстрату ТМБ і стоп-розчину, як описано в попередньому пункті. Коефіцієнт розведення кожного кон’югату, що дає OD450 в межах 1,0-1,5, використовували як робоче розведення для основного експерименту з перехресного інгібування наступним чином.

П'ятдесят мікролітрів робочого розведення кожного HRP-кон’югованого мАТ вносили в мікролунки з антигенним покриттям одночасно з рівним об'ємом ІФА-буфера (референтні лунки) або всі мАТ розводили до 10 мкг/мл в тому ж буфері. Після 30 хвилин інкубації при 37С реакцію завершували, як описано вище. Всі комбінації проводили в двох екземплярах. Дані для кожної комбінації мічених і немічених HRP мАБ представлені у вигляді відсотка інгібування: (середня OD450 фактичної комбінації - середня OD450 референтних лунок)/середня OD450 референтних лунок.

Результати експерименту подані у вигляді таблиці (рисунок 2.1), де кожна комірка це коефіцієнт перехресного зв’язування міченого антитіла (зі стовпця) та не міченого (з рядка). В рядку позначеному як “blank” надані максимальні значення коефіцієнтів перехресного зв’язування для відповідного міченого антитіла.

Зображення, що містить стіл

Автоматично згенерований опис  
Рисунок 2.1 Фрагмент початкових даних

Перед тим як застосовувати той чи інший алгоритм необхідно перетворити дані з таблиці за допомогою формули:

|  |  |
| --- | --- |
|  | (2.1) |

Нові значення представляють відсоткове співвідношення між значенням у клітинці та максимальним значенням для відповідного стовпця. Нові значення знаходяться в діапазоні від 0 до 1.

Основна задача, тепер, полягає у розбитті отриманої матриці перехресного зв’язування антитіл на групи за ознакою подібності показника зв’язування для того, щоб полегшення виявлення оптимальних пар та приблизної локалізації місця зв’язування.

Проте роботу з отриманими даними можна розділити на два типи:

1. Роботу як з категоріальними змінними;
2. Роботу як з кількісними змінними.

## 2.2 Алгоритм роботи з категоріальними змінними

На початковому етапі роботи, для простоти було прийнято, що дані в кожній комірці поділяються на три категорії за наступним алгоритмом:

* 1. цифрою 3 (темно-зеленим кольором) антитіла з поганим зв’язуванням, якщо

|  |  |
| --- | --- |
|  | (2.2) |

* 1. цифрою 2 (світло-зеленим кольором) антитіла з середнім зв’язуванням, якщо

|  |  |
| --- | --- |
|  | (2.3) |

* 1. цифрою 1 (білим кольором) антитіла з хорошим зв’язування – у всіх інших випадках.

Зображення, що містить текст, шафа

Автоматично згенерований опис  
Рисунок 2.2. Фрагмент позначених даних

В даній роботі для категоріальними змінних використовується метод кластеризації k-modes, який є модифікація методу k-середніх.

k-середніх – це популярний алгоритм кластеризації на основі центроїдів, який розділяє дані представлені у вигляді точок на k кластерів, кожен із яких має майже рівну кількість цих точок. Ідея цього алгоритму кластеризації полягає в тому, щоб знайти k центроїд, де кожна точка з набору даних буде належати будь-якій з k-множин з найближчим (найчастіше мінімальна евклідова відстань) до нього середнім значенням.

k-modes метод, розроблений Хуагном в 1998 році, що визначає кластери на основі відповідності категорій між точками даних. В цьому алгоритмі на відміну від k-середніх відстань між двома точками (проста міра не схожості) даних X та Y описується як сума не схожих елементів:

|  |  |
| --- | --- |
|  | (2.4) |

де

|  |  |
| --- | --- |
|  | (2.5) |

Також, середнє змінюється на моду, тобто тепер елемент до кластера будуть відносити не тоді, коли цей кластер має середнє значення, що найближче до цього елемента, а коли мода цього кластера буде найближчою до елемента. Моди відповідно знаходиться на основі частоти появи значення в кластері (приклад наведено в таблиці 2.1).

Таблиця 2.1 Приклад знаходження центроїдів в алгоритмі k-mods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Кластер 1 | Кластер 2 | Кластер 3 | Кластер 4 | Кластер 5 |
| Елемент 1 | 1 | 0 | 1 | 1 | 1 |
| Елемент 2 | 1 | 0 | 0 | 1 | 2 |
| Елемент 3 | 1 | 1 | 0 | 1 | 2 |
| Елемент 4 | 1 | 1 | 0 | 0 | 2 |
| Елемент 5 | 1 | 0 | 0 | 0 | 1 |
| Центроїди | 1 | 0 | 0 | 1 | 2 |

Нижче наведено кроки алгоритму кластеризації на основі k-modes:

1. Вибір k початкових мод.
2. Розділення елементів на кластери на основі простої міри несхожості відносно початкових мод. Після додавання нового елементу відбувається оновлення мод кожного з кластерів.
3. Після того, як усі елементи були віднесені до кластера, відбувається перевірка значення несхожості кожного спостереження з модою. Якщо виявляється, що найближча мода знаходиться в іншому кластері, відбувається переміщення елемента у відповідний кластер і оновлення мод обох кластерів.
4. Повторення кроку 3, доки жоден із елементів не змінить кластер на інший.

Метод k-modes працює лише при початкові відомій кількості кластерів, оскільки вона нам не відома, то для визначення оптимальної кількості кластерів використовується так званий „ліктьовий метод” (elbow method). Даний метод модифікований так, щоб використовувати відмінність всередині кластера (within-cluster difference) як вартість.

|  |  |
| --- | --- |
|  | (2.6) |

де WCD – відмінність всередині кластера, k – кількість кластерів, m – кількість спостережень у кожному кластері, c – центроїд кластера, а – проста міра несхожості.

Суть методу в тому щоб, для різних значень k вибирати значення k у тій точці, де значення істотно не зменшуватиметься зі збільшенням значення k. [3 пункт до рисунку 2.2]

У якості початкового розбиття на кластери використовується алгоритм розроблений Fuyuan Cao в 2009 році. Цей метод ініціалізації для категоріальних даних, у якому враховують відстань між об’єктами та щільність об’єкта, що визначається на основі частоти значень атрибутів. [4]

## 2.3 Алгоритм роботи з кількісними змінними

Оскільки нові значення представляють відсоткове співвідношення між значенням у клітинці та максимальним значенням для відповідного стовпця, то їх модна розглядати як кількісні змінні.

Для кластеризації кількісних змінних було вирішено застосувати один із найпопулярніших видів кластеризації – ієрархічні алгоритми, а саме їх агломеративний підвид. Використання агломеративного підвиду („знизу вгору”) означає, що при побудові кластерів спочатку кожна точка даних належить власному кластеру, і з кожним кроком кластери об’єднуються в більші доки всі точки даних не опиняться в одному кластері. На рисунку 2.3 зображено приклад виконання алгоритму. В результаті уварюється ієрархія, яку зображають з використанням дендрограми (рисунок 2.4).

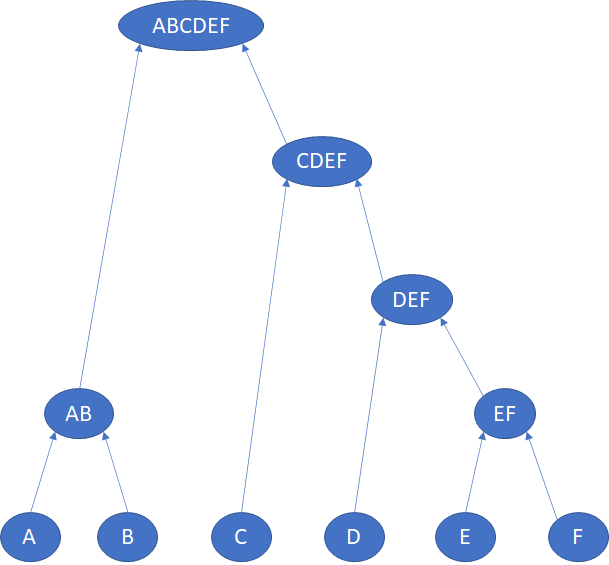


Рисунок 2.3. Приклад агломератної фільтрації

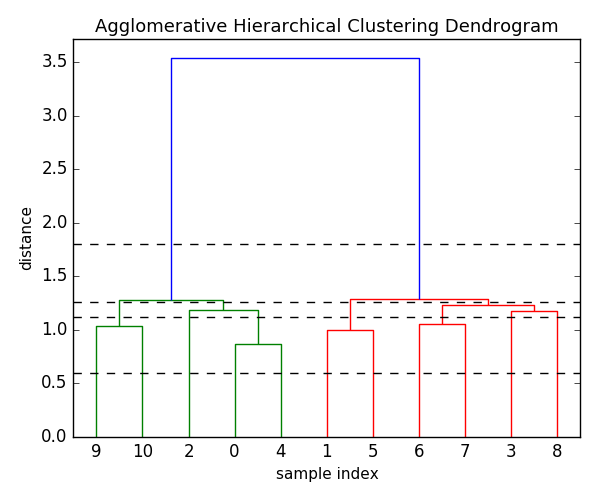


Рисунок 2.4 Приклад дендрограми

Також було обрано декілька стратегій зв’язування[5][дока sas]:

1. Метод Уорда (Ward linkage) – на скільки збільшується дисперсії для кластерів, які об’єднуються;

|  |  |
| --- | --- |
|  | (2.7) |

де – центроїди А та B відповідно.

1. Метод повного зв’язку (Complete linkage) – максимальна відстань між елементами кожного кластера;

|  |  |
| --- | --- |
|  | (2.8) |

1. Метод середнього зв’язку (Average linkage) – середня відстань між елементами кожного кластера;

|  |  |
| --- | --- |
|  | (2.9) |

1. Метод одиночного зв’язку (Single linkage) – мінімальна відстань між елементами кожного кластера;

|  |  |
| --- | --- |
|  | (2.10) |

У всіх вище наведених це метрика на основі якої визначають відстані між кластерами у цій роботі, було вирішено використовувати найпростішу метрику – Евклідову відстань.

|  |  |
| --- | --- |
|  | (2.11) |

# 3 Практична реалізація

В результаті виконання магістерської роботи було розроблено веб-застосунок, який складається з трьох основних частин:

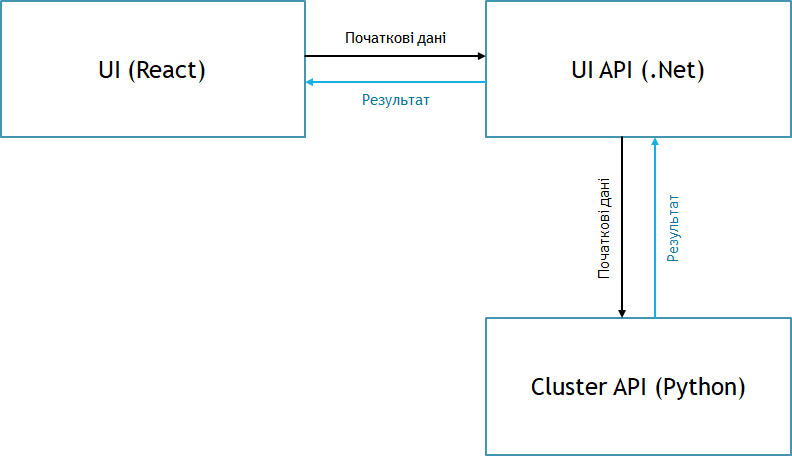
Інтерфейс користувача, написаний на мові програмування TypeScript з використанням бібліотек React та Ant Design, який реалізує функції:

1. Вивантаження файлу з початковими даними
2. Перегляд результатів
3. Завантаження файлу з результатами у форматі .xlsx

Прикладний програмний інтерфейс (API), який відповідає за кластеризацію написаний на мові програмування Python 3. Для завантаження та обробки даних використовувалась бібліотека pandas, для візуалізації бібліотека matplotlib, для кластеризації Scikit-learn, а саме агломеративну кластеризацію зі стратегією зв’язку, що реалізована методом Уорда, та kneed для знаходження оптимальної кількості кластерів.

Також реалізовано прикладний програмний інтерфейс, що поєднує між собою два попередні пункти та реалізує перетворення даних для відображення та збереження результатів. Цей інтерфейс написаний на мові програмування C# з використанням .NET 5 та ASP.NET Core 5.0.

Архітектура веб застосунку зображена на рисунку 3.1.

  
Рисунок 3.1 Архітектура веб застосунку

Користувач завантажує файл з початковими даними, які відправляються на UI API, який в свою чергу відправляє їх на Cluster API, де дані обробляються та відбувається процес кластеризації. Після цього результат відправляється на UI API де конвертується у форму зручну для відображення та віддається на інтерфейс користувача. Повний код UI API та інтерфейсу користувача наведений в додатках А та Б відповідно.

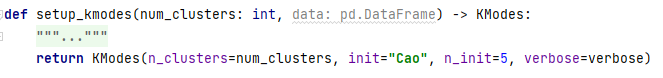
Для експорту результатів дані з UI знову відправляються на UI API, який у свою чергу формує та повертає файл у форматі .xlsx. Для експорту використано бібліотеку з відкритим вихідним кодом, під назвою SpreadsheetLight. Повний код функції для експорту файлу наведений в Додатку В.

Також було реалізовано програму мовою Python для тестування та порівнння різних методів кластеризації. Вона складається з двох частин, перша відповідає за кластеризацію методом k-modes.

Основна функція цієї частини приймає на вхід дані для обрахунку та складається з таких складових:

1. Кластеризація на все можливу кількість кластерів (від 1 до кількості точок даних)
2. Знаходження оптимальної кількості кластерів
3. Кластеризація для оптимальної кількості кластерів

Для кластеризації використовувалась бібліотека kmodes та відповідний метод KModes з неї (фрагмент програми наведено на рисунках 3.2 та 3.3).

  
Рисунок 3.2 Створення класу KModes

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.3 Функція кластеризації з заданою кількістю кластерів

Після того як проведено кластеризацію оптимальна кількість кластерів (фрагмент програми наведено на рисунки 3.4 та 3.5), зважаючи на яку відбувається повторне розбиття на кластери (рисунок 3.3). Для цього використовується KneeLocator з бібліотеки kneed.

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.4 Ліктьовий метод

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.5 Функція знаходження оптимальної кількості кластерів

Основна функція hierarchical, яка приймає на вхід дані для обрахунку, стратегію зв’язування, інтерполяційни метод для знаходження оптимального порогу відстані та підпис до графіка (рисунок 3.6). Також на рисунку 3.6 зображено приклад виклику функції для стратегії зв’язування за методом Уорда.

Функція hierarchical складається з 3-ьох основних частин:

1. Кластеризація
2. Знаходження оптимального порогу відстань
3. Побудова дендрограми

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.6 Головна функція ієрархічної кластеризації

Для кластеризації використовувалась бібліотека Scikit-learn та метод AgglomerativeClustering з неї який приймає на вхід стратегію зв’язування, поріг відстані та інші параметри (фрагмент програми наведено на рисунку 3.7).

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.7 Функція кластеризації

Після того як проведено кластеризацію знаходиться поріг відстані (фрагмент програми наведено на рисунку 3.8), зважаючи на яку відбувається розбиття на кластери. Для цього використовується KneeLocator з бібліотеки kneed. Який використовує один з двох інтерполяційних методів:

1. interp1d – використовує метод „scipy.interpolate.interp1d”.
2. polynomial – використовує метод „[numpy.polyfit](https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html)”, та призначається для гладкіших кривих.

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.8 Функція знаходження оптимального порогу відстані

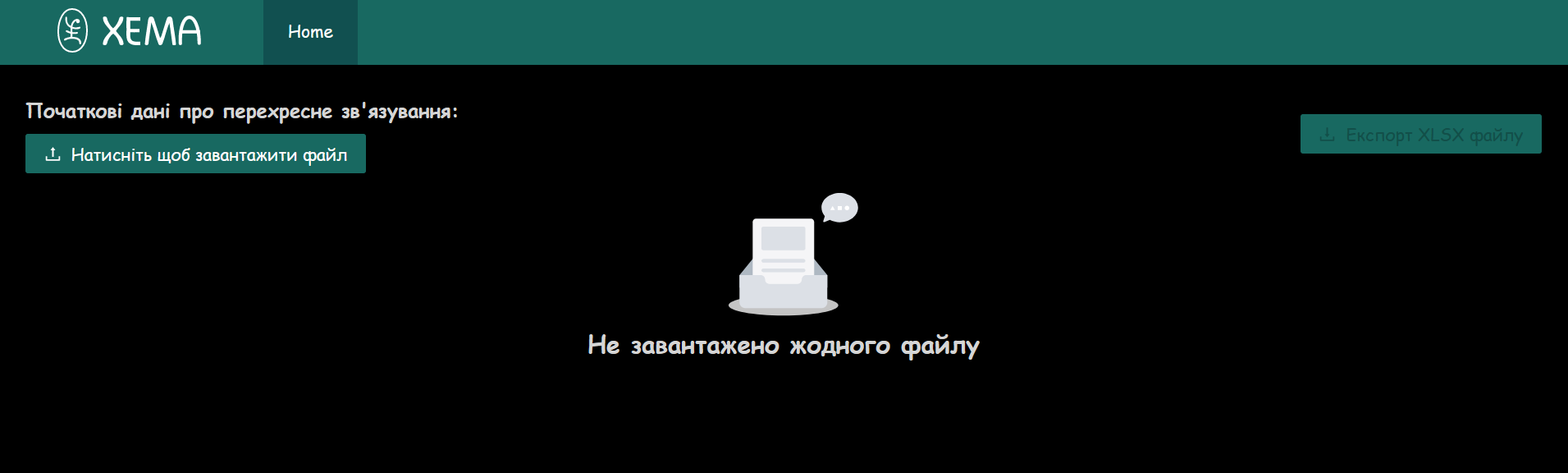
Після того, як поріг відстані знайдено будується матриця зв’язку та на її основі малюється дендрограма (фрагмент програми наведено на рисунку 3.9).

Зображення, що містить текст

Автоматично згенерований опис  
Рисунок 3.9 Функція для побудови дендрограми

# 4 Результати

Розроблений веб застосунок має декілька сторінок на рисунку 4.1 зображено початкову сторінку на якій можна завантажити файл з даними про перехресне зв’язування антитіл.

  
Рисунок 4.1 Початкова сторінка веб застосунку

Результуючі дані на інтерфейсі користувача дані відображаються у два способи:

1. У вигляді карток де заголовок вказує на групу, а в тілі міститься посортований список антитіл (рисунок 4.2).
2. У вигляді початкових даних, позначених кольорами (рисунок 4.3). Співпадіння кольорів в тій чи іншій колонці в середині певного кластеру підтверджує, що вони справді належать до однієї групи.

Зображення, що містить текст, знімок екрана, екран, срібло

Автоматично згенерований опис  
Рисунок 4.2 Результат кластеризації у вигляді карток

  
Рисунок 4.3 Фрагмент початкових даних розбитих на групи

Також реалізовано збереження результатів у файл формату .xlsx. Ці результати також подані у вигляді таблиці та схожі на результати подані на веб сторінці і складаються з:

1. Таблиці де заголовок вказує на групу, а в тілі міститься посортований список антитіл (рисунок 4.4).
2. Таблиця з початковими даними розбита на групи та позначена кольорами (рисунок 4.5).

Зображення, що містить стіл

Автоматично згенерований опис  
Рисунок 4.4 Результат кластеризації записаний у файл

Зображення, що містить стіл

Автоматично згенерований опис  
Рисунок 4.5 Фрагмент початкових даних розбитих на групи та записаних у файл

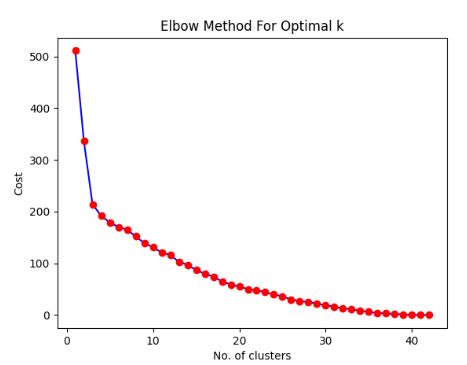
Тепер перейдемо безпосередньо до аналізу числових результатів. Очікувалось, що початкові дані будуть по груповані так як зображено в таблиці 4.1.

Таблиця 4.1 Очікуваний результат кластеризації

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1A | 1B | 1B/2 | 2 | 2B/3 | 3A | 3B | 4A | 4B | 4C | 5 |
| NP1501 | X190 | NP1512 | NP1502 | NP1528 | X202 | X32 | NP3706 | X211 | X215 | X220 |
| NP1514 | NP1526 | NP1521 | NP1503 |  | X218 | X41 |  |  |  | X275 |
| NP1516 | X200 |  | NP1508 |  | NP1518 | X155 |  |  |  |  |
| NP1517 | X201 |  | NP1510 |  | NP1527 | X212 |  |  |  |  |
| NP1507 |  |  | NP1520 |  |  | X213 |  |  |  |  |
|  |  |  | NP1522 |  |  | X217 |  |  |  |  |
|  |  |  | NP1525 |  |  | X223 |  |  |  |  |
|  |  |  | X221 |  |  | X224 |  |  |  |  |
|  |  |  | X271 |  |  | X233 |  |  |  |  |
|  |  |  | NP3701 |  |  | NP1524 |  |  |  |  |
|  |  |  | NP3708 |  |  | NP3715 |  |  |  |  |

## 4.1 Результати роботи алгоритм для категоріальних змінних

Спочатку знаходиться WCD для кожного k – кількість кластерів від 0 до розміру вибірки, та на основі цих даних будується графік та з використанням ліктьового методу знаходять оптимальна кількість кластерів. Для наших даних з графіку WCD до кількості кластерів (рисунок 4.6) видно, що оптимальна кількість кластерів буде 9.

  
Рисунок 4.6 Графік кількості кластерів до відстані в середині кластера

Коли оптимальна кількість кластерів знайдена, відбувається безпосередня кластеризація методом k-modes. Результат кластеризації наведено в таблиці 4.2.

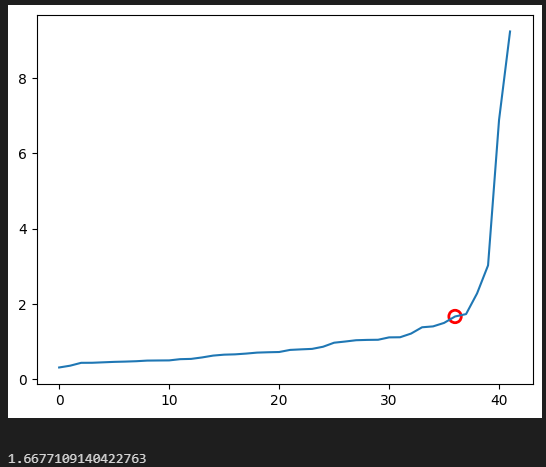
Таблиця 4.2 Результат кластеризації методом k-modes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| NP1501 | NP1528 | NP1502 | NP1524 | NP1526 | NP1507 | X155 | NP1518 | NP1527 |
| NP1512 |  | NP1503 | NP3715 | NP3701 |  | X202 |  | NP3706 |
| NP1514 |  | NP1508 | X212 | X200 |  | X213 |  |  |
| NP1516 |  | NP1510 | X217 | X201 |  | X233 |  |  |
| NP1517 |  | NP1520 | X223 |  |  | X32 |  |  |
| NP1521 |  | NP1522 | X224 |  |  | X41 |  |  |
| X211 |  | NP1525 |  |  |  |  |  |  |
| X215 |  | NP3708 |  |  |  |  |  |  |
| X218 |  | X190 |  |  |  |  |  |  |
| X220 |  | X221 |  |  |  |  |  |  |
| X275 |  | X271 |  |  |  |  |  |  |

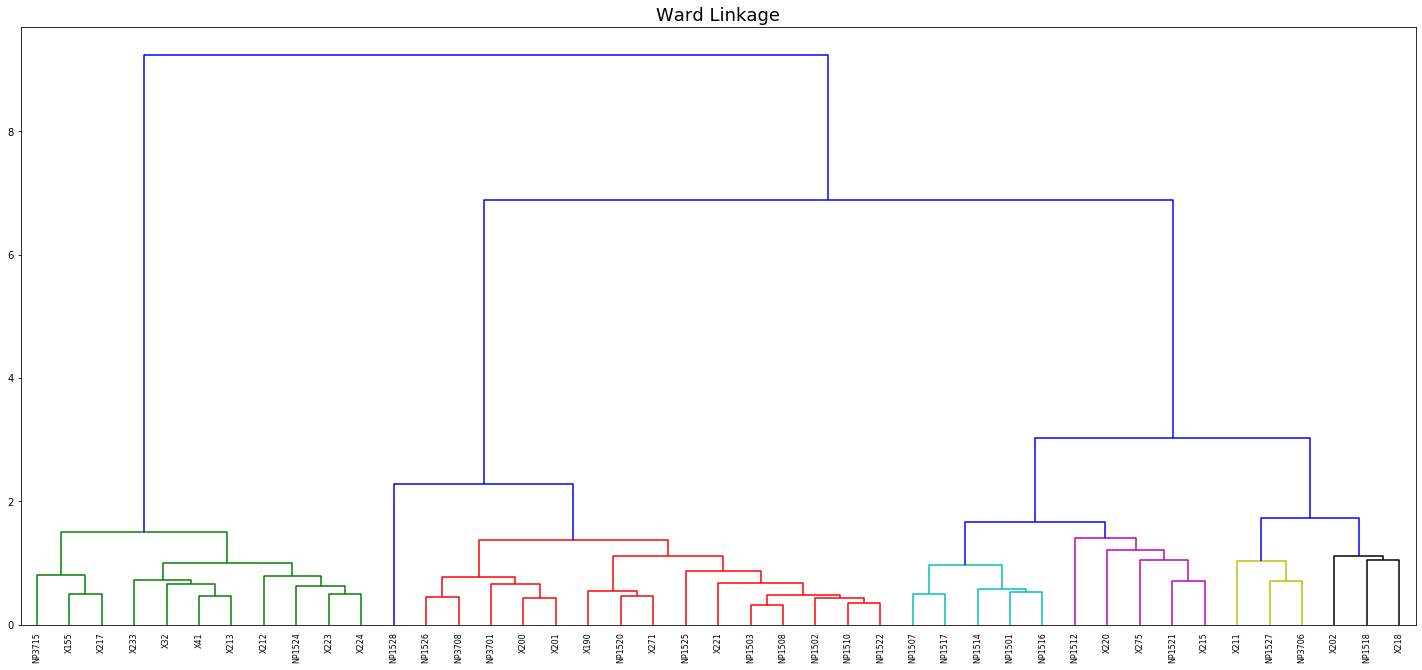
Як можна побачити з таблиці в методі k-modes лише один кластер 2 повністю співпадає з групою 2B/3. Кластер 3 майже повторює групу 2 за винятком елемента NP 3701. Кластер 4 та кластер 7 містять в собі групу 3B та елемент X202.

## 4.1 Результати роботи алгоритм для кількісних змінних

В результаті роботи програми ми отримали 4 виводи для кожної стратегії зв’язування. Спершу розглянемо результати використання метода Уорда. Для нього оптимальний поріг відстані буде рівним 1.67, що видно з рисунку 4.7.

  
Рисунок 4.7 Графік відстаней для методу Уорда

На дендрограмі (рисунок 4.8) чітко видно, що в результаті роботи алгоритму було виділено 7 кластерів. У таблиці 4.3 наведено результат кластеризації, де кожна колонка містить перелік антитіл, які належать до кластеру.

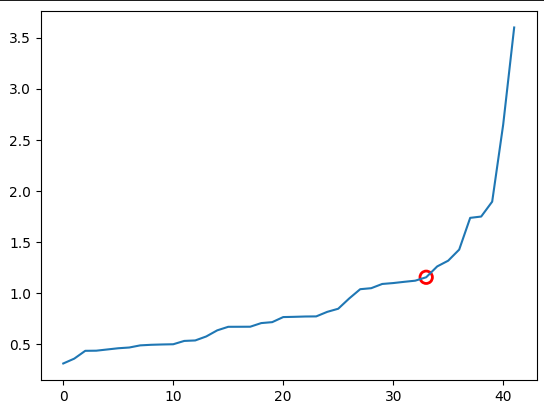
Рисунок 4.8 Дендрограма кластеризації з використанням методу Уорда

Чітко видно, що кластер під номером 1 відповідає групі 3В, кластер 5 повністю відповідає групі 2В/3 і кластер 7 відповідає групі 1А (вони позначені зеленим кольором). Також кластер 3 об’єднує групи 1В і 2, кластер 4 відповідає групі 3А без елемента NP1527, який знаходиться в кластері 6, що також містить групи 4А і 4В (вони позначені жовтим і помаранчевим кольорами).

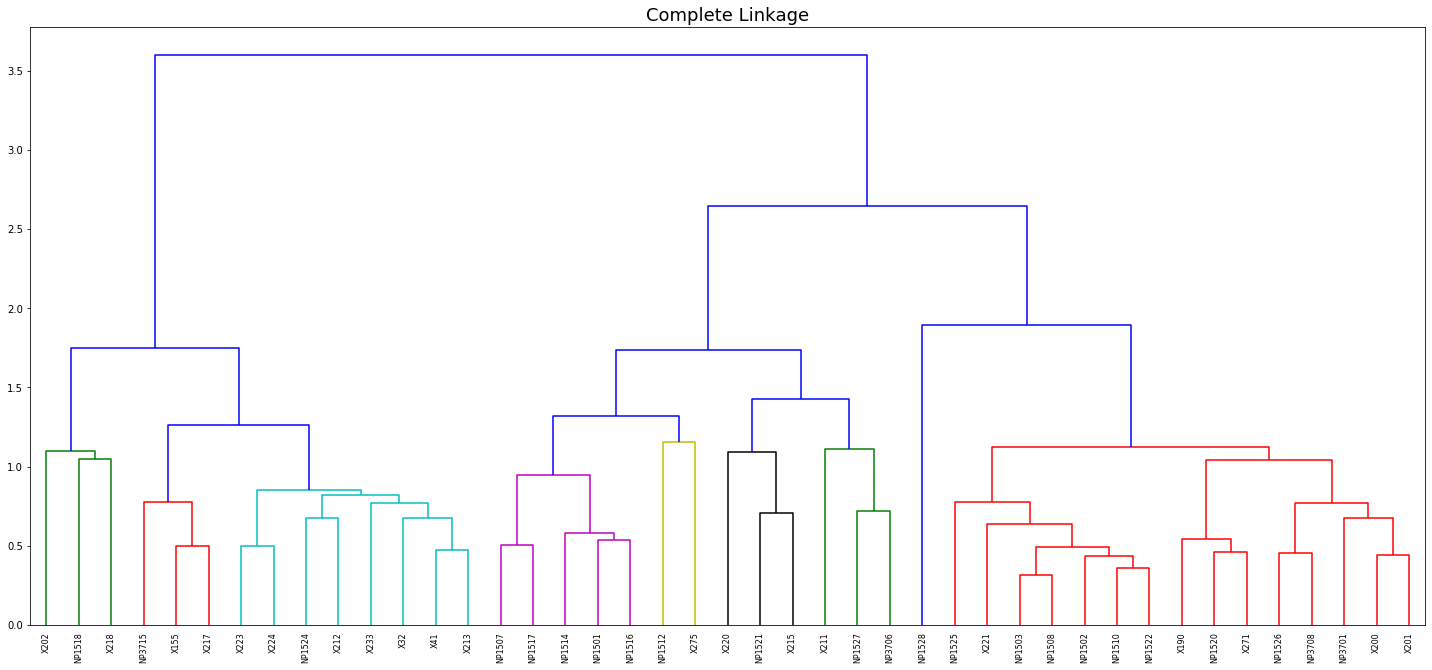
Таблиця 4.3 Результат кластеризації з використанням методу Уорда

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| X41 | NP1521 | X200 | X202 | NP1528 | NP1527 | NP1517 |
| X32 | X275 | X190 | NP1518 |  | X211 | NP1514 |
| X233 | X215 | X221 | X218 |  | NP3706 | NP1516 |
| X224 | NP1512 | X201 |  |  |  | NP1507 |
| X223 | X220 | X271 |  |  |  | NP1501 |
| X217 |  | NP3708 |  |  |  |  |
| X213 |  | NP3701 |  |  |  |  |
| X212 |  | NP1526 |  |  |  |  |
| X155 |  | NP1525 |  |  |  |  |
| NP3715 |  | NP1522 |  |  |  |  |
| NP1524 |  | NP1520 |  |  |  |  |
|  |  | NP1502 |  |  |  |  |
|  |  | NP1503 |  |  |  |  |
|  |  | NP1510 |  |  |  |  |
|  |  | NP1508 |  |  |  |  |

Після цього розглянемо результати використання методу повного зв’язку. Для нього оптимальний поріг відстані буде рівним 1.16, що видно з рисунку 4.9.

  
Рисунок 4.9 Графік відстаней для методу повного зв’язку

На дендрограмі (рисунок 4.10) чітко видно, що в результаті роботи алгоритму було виділено 9 кластерів. У таблиці 4.4 наведено результат кластеризації.

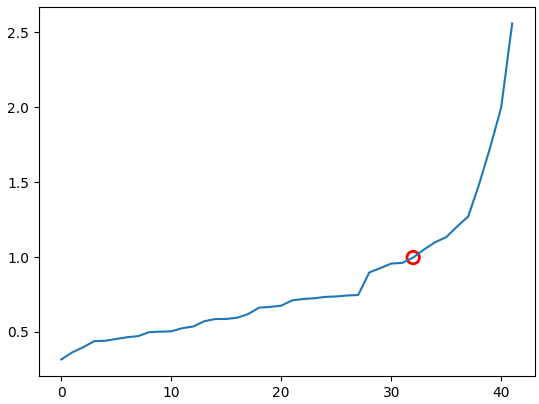
Рисунок 4.10 Дендрограма кластеризації з використанням методу повного зв’язку

Таблиця 4.4 Результат кластеризації з використанням методу повного зв’язку

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| X275 | X221 | X211 | NP1507 | X202 | X213 | NP1521 | NP1528 | X217 |
| NP1512 | X201 | NP1527 | NP1501 | NP1518 | X32 | X215 |  | X155 |
|  | X200 | NP3706 | NP1516 | X218 | X41 | X220 |  | NP3715 |
|  | X190 |  | NP1517 |  | X233 |  |  |  |
|  | X271 |  | NP1514 |  | X224 |  |  |  |
|  | NP3701 |  |  |  | NP1524 |  |  |  |
|  | NP1526 |  |  |  | X212 |  |  |  |
|  | NP1525 |  |  |  | X223 |  |  |  |
|  | NP1522 |  |  |  |  |  |  |  |
|  | NP3708 |  |  |  |  |  |  |  |
|  | NP1502 |  |  |  |  |  |  |  |
|  | NP1503 |  |  |  |  |  |  |  |
|  | NP1520 |  |  |  |  |  |  |  |
|  | NP1508 |  |  |  |  |  |  |  |
|  | NP1510 |  |  |  |  |  |  |  |

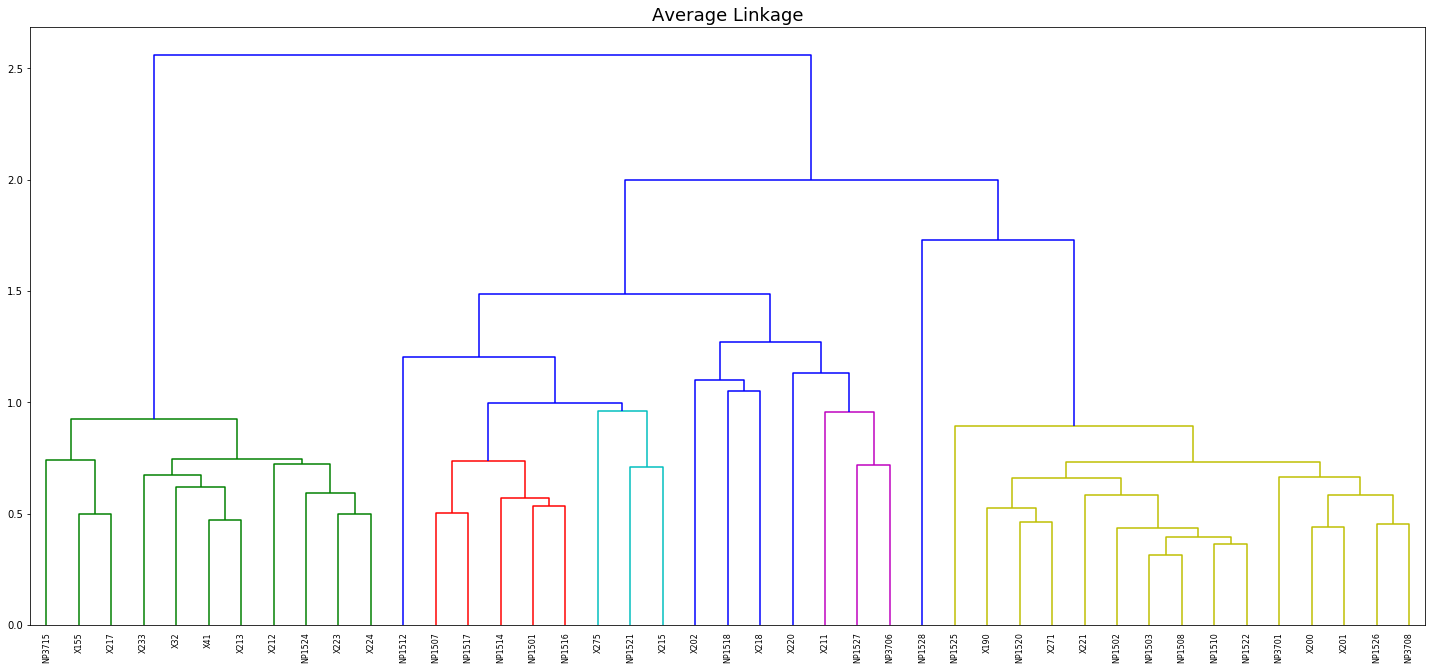
Чітко видно, що кластер номер 4 відповідає групі 1А, а кластер номер 8 повністю відповідає групі 2B/3 (вони позначені зеленим кольором). Також кластер 2 об’єднує групи 1В і 2, кластер 5 відповідає групі 3А без елемента NP1527, який знаходиться в кластері 3, який також містить групи 4А і 4В, крім того кластер 6 і кластер 9 містять елементи з групи 3В (вони виділені жовтим і помаранчевим кольорами).

Третім розглянемо метод середнього зв’язку. Для нього оптимальний поріг відстані буде рівним 0.99, що видно з рисунку 4.11.

  
Рисунок 4.11 Графік відстаней для методу повного зв’язку

На дендрограмі (рисунок 4.12) чітко видно, що в результаті роботи алгоритму було виділено 11 кластерів. У таблиці 4.5 наведено результат кластеризації.

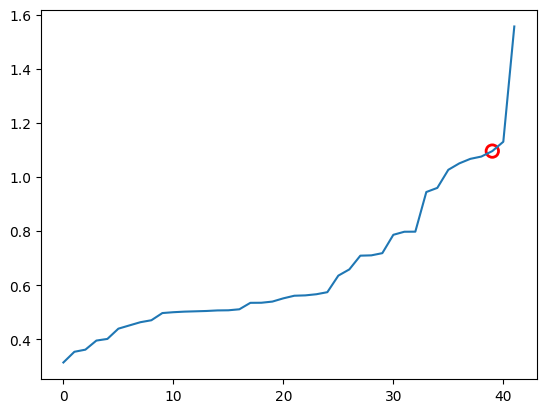
Чітко видно, що кластер номер 2 відповідає групі 3В, кластер 5 повністю відповідає групі 1А і кластер 8 відповідає групі 2В/3 (вони позначені зеленим кольором). Також кластер 6 об’єднує групи 1В і 2, кластери 9, 10 і 11 відповідають групі 3А без елемента NP1527, який знаходиться в кластері 6, який також містить групи 4А і 4В (вони позначені жовтим і помаранчевим кольорами).

Рисунок 4.12 Графік відстаней для методу повного зв’язку

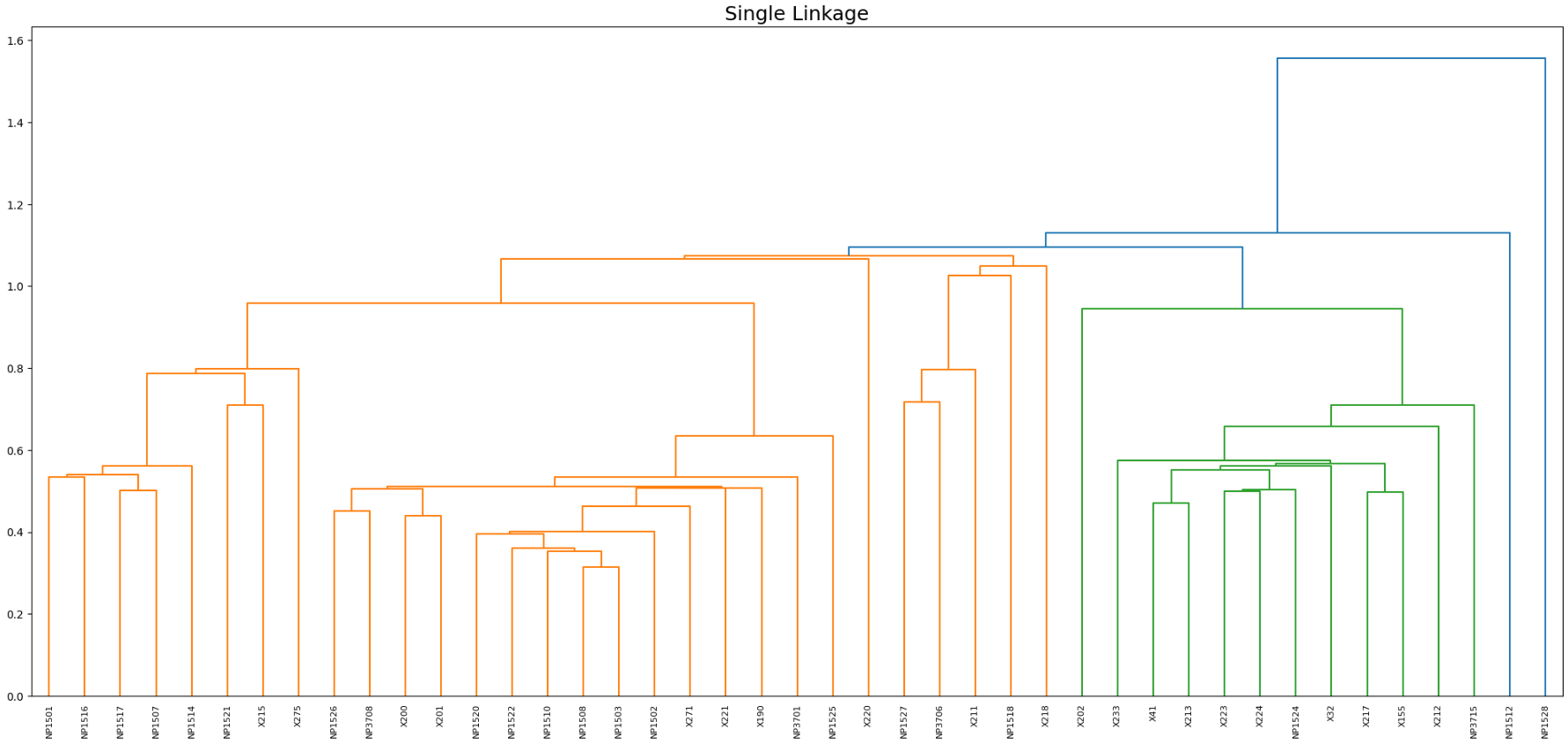
Таблиця 4.5 Результат кластеризації з використанням методу повного зв’язку

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9, 10, 11 | | |
| X275 | NP3715 | X211 | X220 | NP1501 | NP1502 | NP1512 | NP1528 | X202 | NP1518 | X218 |
| NP1521 | NP1524 | NP3706 |  | NP1517 | NP1503 |  |  |  |  |  |
| X215 | X32 | NP1527 |  | NP1507 | X221 |  |  |  |  |  |
|  | X41 |  |  | NP1514 | NP1508 |  |  |  |  |  |
|  | X155 |  |  | NP1516 | NP1510 |  |  |  |  |  |
|  | X212 |  |  |  | NP3701 |  |  |  |  |  |
|  | X213 |  |  |  | X200 |  |  |  |  |  |
|  | X217 |  |  |  | X190 |  |  |  |  |  |
|  | X223 |  |  |  | NP1520 |  |  |  |  |  |
|  | X224 |  |  |  | NP1522 |  |  |  |  |  |
|  | X233 |  |  |  | NP1525 |  |  |  |  |  |
|  |  |  |  |  | X271 |  |  |  |  |  |
|  |  |  |  |  | NP1526 |  |  |  |  |  |
|  |  |  |  |  | X201 |  |  |  |  |  |
|  |  |  |  |  | NP3708 |  |  |  |  |  |

Останнім розглянемо метод одиночного зв’язку. Для нього оптимальний поріг відстані буде рівним 1.095, що видно з рисунку 4.13.

  
Рисунок 4.13 Графік відстаней для методу одиничного зв’язку

На дендрограмі (рисунок 4.14) чітко видно, що на виході алгоритму було виділено 4 кластери. У таблиці 4.6 наведено результат кластеризації.

  
Рисунок 4.14 Графік відстаней для методу одиничного зв’язку

З розподілу даних очевидно, що метод одиничного зв’язку отримав завелике порогове значення, тому що кластер під номером 1 об’єднав в собі майже всі групи. Кластер номер 2 складається з групи 3B та елемента X202, а кластер 8 відповідає групі 2В/3.

Таблиця 4.6 Результат кластеризації з використанням методу одиничного зв’язку

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
| NP1527 | X213 | NP1512 | NP1528 |
| NP3706 | X32 |  |  |
| X201 | X41 |  |  |
| X200 | X155 |  |  |
| X221 | NP1524 |  |  |
| X190 | X212 |  |  |
| NP3708 | NP3715 |  |  |
| NP3701 | X217 |  |  |
| NP1502 | X223 |  |  |
| NP1526 | X224 |  |  |
| NP1525 | X233 |  |  |
| NP1522 | X202 |  |  |
| NP1503 |  |  |  |
| NP1508 |  |  |  |
| NP1520 |  |  |  |
| X271 |  |  |  |
| NP1510 |  |  |  |
| NP1517 |  |  |  |
| NP1514 |  |  |  |
| NP1507 |  |  |  |
| NP1516 |  |  |  |
| NP1501 |  |  |  |
| X220 |  |  |  |
| X218 |  |  |  |
| NP1521 |  |  |  |
| X215 |  |  |  |
| NP1518 |  |  |  |
| X275 |  |  |  |
| X211 |  |  |  |

Отже, якщо за метрику точності взяти сумарну кількість елементів у кластерах, які повністю відповідають очікуваному результату, то очевидно, що найкращий результат дає алгоритм, який використовував метод Уорда та метод повного зв’язку. Алгоритм, який використовував метод середнього зв’язку дав не набагато гірший результат. А алгоритм одиночного зв’язку дав найгірший результат зважаючи на те що він об’єднав багато груп в один кластер.

# Висновок

В результаті виконання магістерської було розроблено веб застосунок, який полегшує задачу розбиття списку антитіл на групи на основі індексу перехресного зв’язування. Цей застосунок представляє з себе три мікросервіси:

1. Інтерфейс користувача для вивантаження файлів, перегляду та завантаження результатів.
2. Прикладний програмний інтерфейс, який відповідає за кластеризацію
3. API, що поєднує інтерфейс для кластеризацію та інтерфейс користувача, а також реалізує завантаження результатів.

Для розробки системи було використано NET 5 з ASP.NET, Python 3 з Flask та Type Script з React і Ant Design.

Також було виконано кластеризацію антитіл для конкретного вірусного протеїну. Зважаючи на те, які антитіла опинились в однакових групах порівнюючи очікуваний та отриманий результат можна сказати, що правильність розбиття на кластери склала 70%.

Важливим аспектом, є те, що попри невеликий об’єм даних (матриця 40х30) розроблений застосунок за невеликий час (1-2 хвилини) робить об’єм роботи, для виконання якого людині потрібно було б декілька днів. Це показує, що даний застосунок є корисним, оскільки при збільшенні вибірки кількість часу, що потрібен на виконання алгоритму комп’ютером буде залишатись невеликим в порівнянні з часом який потрібен буде людині для виконання такого ж завдання.

# Список Використонаї Літератури

1. Зелінський О. Розробка системи кластеризації антитіл на основі коефіцієнту перехресного зв’язування / О. Зелінський, В. Горлач, Ю. Лебедін // Міжнародна студентська наукова конференція з питань прикладної математики та комп’ютерних наук (МСНКПМК-2022), 5-6 травня 2022 р. – Львів:2022. – С. 8-12. Режим доступу: https://ami.lnu.edu.ua/wp-content/uploads/2022/05/ISSCAMCS-2022.pdf
2. Z. Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values (1998). Data Mining and Knowledge Discovery. 2(3): 283–304.
3. Audhi Aprilliant The k-modes as Clustering Algorithm for Categorical Data Type [Electronic resource]. – 2021. – URL: https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7.
4. Fuyuan Cao A new initialization method for categorical data clustering, Fuyuan Cao, Jiye Liang, Liang Bai [Electronic resource]. – 2009. – URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.8181&rep=rep1&type=pdf
5. F. Nielsen, Introduction to HPC with MPI for Data Science, Chapter 8: Hierarchical Clustering, Springer, Switzerland, 195–211. [Electronic resource]. – 2016. – URL: https://www.researchgate.net/publication/314700681\_Hierarchical\_ Clustering. <http://dx.doi.org/10.1007/978-3-319-21903-5_8>.
6. [книжка з біології 9 клас 217-218 cт] Біологія: підручник для 9класу загальноосвітніх навчальних закладів / В. І. Соболь. – Кам’янець-Подільський: Абеткаб 2017ю – 288с.
7. [енциклопедія] Скок М. В. Антитіло // Велика українська енциклопедія. URL: https://vue.gov.ua/Антитіло (дата звернення: 6.11.2022). https://pidruchnyk.com.ua/912-biologiya-sobol-9-klas.html
8. [словник] Словник української мови: в 11 томах. — Том 1, 1970. — Стор. 50. <http://sum.in.ua/s/antytila>
9. [методичка з екології] <https://ela.kpi.ua/bitstream/123456789/15376/1/Klasternyi_analis.pdf>
10. [документація sas] <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_cluster_sect012.htm>