Oświadczenie kierującego pracą

Oświadczam, że praca dyplomowa magisterska studenta Pauliny Czubaj pt. „Algorytmy i metody klasyfikacji obiektów 2D z zastosowaniem do analizy artefaktów dziedzictwa kulturowego.” została przygotowana pod moim kierunkiem, stwierdzam, że spełnia ona warunki przedstawienia jej w postępowaniu o nadanie tytułu zawodowego magistra.

Data .....

.....

podpis kierującego pracą

Oświadczenie autora pracy

Świadomy odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa magisterska pt. „Algorytmy i metody klasyfikacji obiektów 2D z zastosowaniem do analizy artefaktów dziedzictwa kulturowego.” została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami (Ustawa z dnia 04.02.1994 r. o prawie autorskim i prawach pokrewnych (tekst jednolity: Dz. U. z 2006 r. nr 90, poz. 631 z późniejszymi zmianami).

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w szkole wyższej.

Oświadczam ponadto, że niniejsza wersja jest identyczna z załączoną wersją elektroniczną umieszczoną w systemie APD.

Data .....

.....

podpis autora pracy

UNIWERSYTET JAGIELLOŃSKI

WYDZIAŁ MATEMATYKI I INFORMATYKI

Kierunek Matematyka Komputerowa

Paulina Maria Czubaj

ALGORYTMY I METODY KLASYFIKACJI OBIEKTÓW 2D Z ZASTOSOWANIEM DO ANALIZY ARTEFAKTÓW DZIEDZICTWA KULTUROWEGO

praca magisterska napisana
w Katedrze Matematyki Obliczeniowej
pod kierunkiem
dr. Marcina Żelawskiego

Kraków, 2019

Algorithms and methods for clustering of 2D contours with application to cultural heritage artifacts analysis

Słowa kluczowe:

- ✓ klastrowanie
- ✓ miary podobieństwa
- ✓ analiza skupień
- ✓ rozpoznawanie i przetwarzanie obrazów

Keywords:

- ✓ clustering
- ✓ similarity measures
- ✓ cluster analysis
- ✓ image recognition and processing

Niniejszą pracę pragnę zadedykować moim rodzicom,
którzy dali mi możliwość kształcenia się i wierzyli we mnie na każdym jego etapie.
Mamie Danucie, za to, że nauczyła mnie uporu w dążeniu do stawianych sobie celów,
tacie Sławomirowi, za ukazanie piękna świata nauki i inspiracje w pracy naukowej.
Pracę pragnę zadedykować również mojemu narzeczonemu, Mariuszowi,
za wiarę we mnie, dodawanie otuchy oraz wsparcie.

Serdeczne podziękowania za cenne rady, cierpliwość i wyrozumiałość
okazaną mi podczas pisania mojej pracy składam na ręce
dr. Marcina Żelawskiego.
Pragnę również podziękować za współpracę merytoryczną oraz nieocenione wskazówki
dotyczące zagadnień poruszanych w pracy
prof. Ewie Bednarczuk, dr. Monice Sydze oraz mgr. Agnieszce Kaliszewskiej.
Praca ta powstała we współpracy pomiędzy Wydziałem Matematyki i Informatyki Uniwersytetu
Jagiellońskiego a Instytutem Badań Systemowych Polskiej Akademii Nauk.

Abstrakt

Celem niniejszej pracy jest dokonanie przeglądu i analizy metod i algorytmów klasyfikacji obiektów dwuwymiarowych. Metody te zostały zewaluowane pod kątem skuteczności ich działania na zbiorach artefaktów dziedzictwa kulturowego reprezentowanych przez zbiory ceramik pochodzenia wykopaliskowego. W pracy zastosowane zostały znane metody i algorytmy oraz podejście autorskie. Efektem pracy jest stworzenie narzędzia do automatycznego grupowania obiektów.

Abstract

The aim of this thesis is to review and analyze the methods and algorithms for classifying two-dimensional objects. These methods have been evaluated in terms of their effectiveness on the collections of cultural heritage artefacts represented by collections of ceramics of archaeological origin. The thesis uses well-known methods and algorithms created by the author. The effect of this thesis is to create a tool for automatic grouping of objects.

Spis treści

Wstęp

W ostatnich latach dynamicznie rozwija się poszukiwanie rozwiązań klasyfikacji wielkoformatowych oraz licznych zbiorów danych, tzw. big data. Najbardziej popularnymi obecnie metodami, stosowanymi do tego typu problemów, są metody oparte o zagadnienia sztucznej inteligencji. Problem pojawia się jednak w momencie, gdy analizowane zbiory okazują się być małoliczne. W takich przypadkach niezbędne jest odejście od podejścia uczenia maszynowego do skupienia się na analizie podobieństw obiektów podobnych bądź różnic obiektów różnych.

Przykładem zbiorów małolicznych są zbiory fragmentów ceramicznych pochodzących z wykopalisk archeologicznych. Zbiory te są podstawową kategorią odkryć dokonywanych podczas prac archeologów [13]. Podstawowymi cechami tych zbiorów są ich nieregularność, niedokładność oraz zachodzące procesy korozyjne, co czyni zastosowanie analizy 3D do klasyfikacji zabiegiem nieefektywnym [1]. Stąd zachodzi potrzeba znalezienia innego podejścia do analizy tychże ceramik. Również cechą charakterystyczną zbiorów wykopalisk jest mała liczność dostępnych danych, do których rozwiązania mają być stosowane. Celem przeprowadzonych badań jest więc znalezienie metod bazujących na podobieństwach i różnicach między analizowanymi fragmentami, tak aby metody były skuteczne nawet na małych zbiorach i uniwersalne w swoim zastosowaniu (nie wymagające wielu reprezentantów danej klasy). Wynikiem przeprowadzonych badań i analiz jest narzędzie stworzone i dedykowane do automatycznej klasyfikacji artefaktów ceramicznych.

Cel i zakres pracy

Celem pracy magisterskiej o powyższym tytule jest analiza problemu klasyfikacji małolicznych zbiorów obiektów 2D oraz stworzenie narzędzia do automatycznej klasyfikacji artefaktów ceramicznych. Obiekty, opisane za pomocą krzywych, są przekrojami naczyń lub ich fragmentów odkrytych w trakcie prac archeologicznych. W trakcie analizy zostaną zbadane dostępne miary podobieństwa dla różnych reprezentacji podanych krzywych, jak również dobór najlepszych metod klastrowania dla tych danych. Zostaną również zaproponowane nowe metody rozwiązania powyższego problemu.

Globalne cele, które zostały wyznaczone podczas formułowania problemu postawionego w powyższej pracy to skrócenie czasu klasyfikacji artefaktów ceramicznych oraz dążenie do otrzymania narzędzia uniwersalnego, co skutkować będzie wysoką efektywnością algorytmu i poprawnością wyników. Spełnienie tych celów, oprócz opracowania wsparcia do użycia podczas prac archeologicznych, mogą również służyć jako wskazówka podczas szkoleń z ręcznego podejścia do klasyfikacji artefaktów ceramicznych [12].

Układ pracy jest następujący. W rozdziale 1. przedstawiono charakterystykę i genezę zbioru testowego oraz zagadnienia teoretyczne związane z analizą krzywych 2D. W rozdziale 2. wprowadzono niezbędne definicje i pojęcia, opisano również metryki i miary wykorzystywane w pracy. Znajdują się tam także teoretyczne zagadnienia zastosowanych w pracy metod. Rozdział 3. poświęcony jest procesowi wstępnej obróbki zbioru wybranego do analizy. W rozdziale 4. zaprezentowano wyniki zastosowania analizy skupień, natomiast w rozdziale 5. przeprowadzono test zastosowania do problemu podejścia syntaktycznego. Wyniki te zostały wygenerowane poprzez autorskie narzędzia służące do automatycznej klasyfikacji artefaktów ceramicznych. W rozdziale 6. opisano inne możliwe podejście do problemu automatycznego grupowania profili. Ostatnią częścią pracy jest ogólne podsumowanie zawierające ocenę jakości uzyskanych rozwiązań i możliwych perspektyw ich rozwoju.

Rozdział 1.
Rys historyczny i archeologiczny problemu

W tym rozdziale znajduje się geneza zbioru oraz zdefiniowanie problemu rozważanego w pracy.

Słownik PWN podaje definicję archeologii - "[gr. archḗ'zasada', 'podstawa', 'starożytny', 'dawny', lógos 'słowo', 'nauka'], nauka historyczna badająca przeszłość dawnych społeczeństw (głównie na podstawie wykopalisk), źródłoznawcza dziedzina szeroko pojmowanej historii (łącznie z prahistorią)" [2]. Na podstawie tej definicji okazuje się więc, że badanie etymologii i genezy wykopalisk jest podstawą dziedziny jaką jest archeologia. Z odkrytych artefaktów ceramicznych badacze są w stanie wyciągnąć wnioski dotyczące m.in. okresu ich wytworzenia, rodzaju wykorzystanych materiałów oraz sposobu wykonania, co prowadzi do kolejnych wniosków na temat życia ludzi w danym okresie czasu i na danej szerokości geograficznej. Okazuje się bowiem, że kształt i geneza każdego artefaktu ma ścisły związek z miejscem i czasem jego pochodzenia [14]. Dzięki temu możliwe są porównania między różnymi regionami świata, podkreślając w ten sposób różnice i podobieństwa w gospodarce okresów historycznych [12].

Najczęściej odkrywanymi rodzajami dowodów archeologicznych, które można znaleźć podczas badań terenowych, są garncarstwo i ceramika. Archeolodzy specjalizujący się w tych dziedzinach stają w obliczu czasochłonnego zadania analizy dziesiątek tysięcy artefaktów. Dość problematyczną jest sytuacja wykopania zaledwie fragmentu artefaktu. Wtedy pierwszym krokiem w analizie znalezionego fragmentu jest ekstrakcja pełnego profilu, a dopiero kolejnym etapem analizy jest przystąpienie do procesu klasyfikacji schematu ogółu profilu.

Klasyfikacja ceramiki uzyskanej w wyniku wykopalisk jest ważną częścią analizy archeologicznej. Badane zestawy czasami mogą być bardzo duże (do tysięcy), czasem składają się natomiast z zaledwie kilku elementów. Klasyfikacje są tradycyjnie wykonywane ręcznie i opierają się wyłącznie na rysunkach naczyń i ich fragmentów. Takie ręcznie tworzone podziały w dużym stopniu zależą od wiedzy i doświadczenia naukowców, w wyniku czego są podatne na tendencyjność. Są także procesem czasochłonnym, co z kolei potęguje jego kosztowność.

Automatyczna klasyfikacja ceramiki pobudziła wielkie zainteresowanie archeologów i informatyków w ostatnich latach [12], [15] – [17]. Proponowano różne próby przezwyciężenia trudności związanych z klasyfikacją obiektów ceramicznych, wśród których godnym uwagi przykładem jest wykorzystanie profilografu. Można także natrafić na informacje o zastosowaniu

do analizy wykopalisk ceramicznych komputerowego skanowania 3D [3] - [6]. Podczas gdy w ostatnich latach technologia skanowania przedmiotów w trzech wymiarach niezwykle się rozwinęła, nie znalazła ona, niestety, praktycznego zastosowania jako narzędzie towarzyszące pracom archeologicznym. Jedną z głównych przeszkód jest brak niezawodnego i wydajnego algorytmu do wydobywania osi symetrii i późniejszego rysowania reprezentatywnego profilu. Zadanie to nie jest trywialne, ponieważ trzeba pokonać kilka przeszkód m.in. fragmenty zwykle pokrywają raczej niewielką część pełnego obwodu oryginalnego naczynia, a im mniejszy fragment, tym trudniej jest ustalić jego prawidłowe położenie. Pierwotne naczynia zwykle nie są idealnie symetryczne – w skali makroskopowej mogą być lekko zdeformowane lub z wnętrzem i zewnętrznymi powierzchniami, które nie są dokładnie koncentryczne. W skali mikroskopowej powierzchnie starożytnej ceramiki są szorstkie ze względu na techniki produkcji i materiały lub ze względu na wietrzenie i pękanie w czasie, który upłynął między ich produkcją a teraźniejszością. Nieregularności te wystarczają do destabilizacji algorytmów pozycjonujących, które doskonale sprawdzają się na gładkich powierzchniach. Powierzchnia 3D uzyskana ze skanera obejmuje punkty, które nie są częścią oryginalnej powierzchni naczynia. Mogą one raczej należeć do powierzchni pęknięcia, która została wytworzona, gdy pierwotne naczynie zostało złamane, lub do defektów powierzchniowych, a ich obecność w modelu jest uciążliwością, którą należy systematycznie usuwać. Przegląd dotychczasowych wniosków na temat zastosowania analizy 3D można znaleźć w [1], [4] – [10]. W przeszłych badaniach znajdujemy również wiele podejść do klasyfikacji tychże obiektów z pominięciem analizy trójwymiarowej. Propozycje rozwiązań postawionego problemu to m.in. analiza środkowych punktów profili fragmentów bądź skupienie się na cechach wizualnych powierzchni fragmentów [17]. Znajdujemy także podejścia do prostej automatyzacji klasyfikacji podstaw ceramicznych [18] oraz rozwiązania bazujące na stworzeniu narzędzi matematycznych do opisu morfologicznego fragmentów artefaktów [10].

Zbiór danych ceramicznych, który został użyty do opracowania algorytmów i metod pochodzi z publikacji P. Mountjoya [19] i przedstawia wybór kilku typów waz późnych helladycznych – pochodzących z kultury mykeńskiej. Wyniki uważa się za dobre, gdy pokrywają się z tradycyjną klasyfikacją (opublikowaną w [19]). W skład zbioru wchodzą krzywe reprezentujące pełne przekroje ceramiczne, nazywane specjalistycznie profilami.

Tradycyjne metody badania wykopalisk ceramicznych opierają się całkowicie na badaniach wykonywanych przez archeologów, oparte są na niedokładnych rysunkach ręcznych i stosunkowo powolnych i drogich badaniach [1]. Rysunki ręczne są rysunkami z kategorii technicznych, czyli rysunków, które są wykonane zgodnie ze ścisłym zestawem ścisłych reguł i są rysowane tak,

aby wizualnie komunikować, jak obiekty są konstruowane - bez konieczności uzupełniania rysunku opisem. W konsekwencji artefakty ceramiczne, reprezentowane jako krzywe, są podawane w standaryzowanej formie, tj. nie są podatne na przypadkowość w ich położeniu w układzie współrzędnych OXY.

Reprezentacja naczyń ceramicznych i ich fragmentów jest reprezentacją czytelną. Zakładając symetrię względem osi OY, krzywe są zarysami przekrojów obiektów z osią OY przyjętą jako oś obrotu. Zdecydowana większość obiektów ma kształt zaokrąglony (gdy jest uformowany na kole garncarskim), co oznacza, że odcinek i położenie osi obrotu są wystarczające, aby przekazać kształt naczynia. Odległość między osią obrotu a najbardziej oddalonym od niej punktem leżącym na krzywej reprezentującej artefakt oznacza promień obiektu. Wszystkie te zabiegi składają się na tezę, że uniwersalizm w reprezentowaniu obiektów skutkować powinien powodzeniem w automatycznym lub półautomatycznym porównaniu rysunków. Warto wspomnieć, że istnieją metody automatycznego generowania profili fragmentów ceramiki ze skanów 3D [9] i zawarte w nich odniesienia. Zbiór testowy rozważany w tej pracy jest zbiorem tworzonym ręcznie.

Rozdział 2.
Zagadnienia teoretyczne

W tym rozdziale wprowadzone zostaną zagadnienia wykorzystywane w pracy. Również, znajduje tu swoje miejsce wykaz najpopularniejszych miar dedykowanych do danych binarnych.

Definicja 2.1. [21] Miary dedykowane do danych binarnych

Niech X, Y – dowolne niepuste ciągi binarne, $|X| = |Y| = n$. Niech d_{xy} – liczba sytuacji, gdzie $x \in X, y \in Y$. Niech więc d_{00}- liczba sytuacji, gdy na tych samych współrzędnych występują wyrazy o wartości 0, d_{11}- liczba sytuacji, gdy na tych samych współrzędnych występują wyrazy o wartości 1, d_{01}- liczba sytuacji, gdy na tych samych współrzędnych występują: wyraz o wartości 0 w ciągu X, wyraz o wartości 1 w ciągu Y, d_{10}- liczba sytuacji, gdy na tych samych współrzędnych występują: wyraz o wartości 1 w ciągu X, wyraz o wartości 0 w ciągu Y.

1. Odległość Hamminga (ang. Hamming distance): $d_{01} + d_{10}$
2. Odległość Euklidesa (ang. Euclidean distance): $\sqrt{d_{01} + d_{10}}$
3. Odległość kwadratowa Euklidesa (ang. Squared-Euclidean distance): $\sqrt{(d_{01} + d_{10})^2}$
4. Odległość średnia taksówkowa (ang. Mean - Manhattan distance): $\frac{d_{01}+d_{10}}{d_{11}+d_{01}+d_{10}+d_{00}}$
5. Odległość Vari'ego (ang. Vari distance): $\frac{d_{01}+d_{10}}{4(d_{11}+d_{01}+d_{10}+d_{00})}$
6. Odległość różnicy rozmiaru (ang. Size - Difference distance): $\frac{(d_{01}+d_{10})^2}{(d_{11}+d_{01}+d_{10}+d_{00})^2}$
7. Odległość różnicy kształtu (ang. Shape - Difference distance): $\frac{n(d_{01}+d_{10})-(d_{01}-d_{10})^2}{(d_{11}+d_{01}+d_{10}+d_{00})^2}$
8. Odległość różnicy wzorca (ang. Pattern - Difference distance):$\frac{4d_{01}d_{10}}{(d_{11}+d_{01}+d_{10}+d_{00})^2}$
9. Odległość Lance'a & Williamsa (ang. Lance & Williams distance): $\frac{d_{01}+d_{10}}{2d_{11}+d_{01}+d_{10}}$
10. Odległość Bray'a & Curtisa (ang. Bray & Curtis distance): $\frac{d_{01}+d_{10}}{2d_{11}+d_{01}+d_{10}}$
11. Odległość Hellingera (ang. Hellinger distance): $2\sqrt{(1 - \frac{d_{11}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})}})}$
12. Odległość cięciwy (ang. Chord distance): $\sqrt{2(1 - \frac{d_{11}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})}})}$
13. Odległość Yule'a (ang. Yule distance): $\frac{2d_{01}d_{10}}{d_{11}d_{00}-d_{01}d_{10}}$
14. Współczynnik podobieństwa Jaccarta - Needhama (ang. Jaccard - Needham similarity): $\frac{d_{11}}{d_{11}+d_{01}+d_{10}}$

15. Współczynnik podobieństwa Dice'a - Sørensena (ang. Dice - Sørensen similarity): $\frac{2d_{11}}{2d_{11}+d_{01}+d_{10}}$
16. Współczynnik podobieństwa 3W - Jaccarta (ang. 3W - Jaccard similarity): $\frac{3d_{11}}{3d_{11}+d_{01}+d_{10}}$
17. Współczynnik podobieństwa Nei'a & Li'a (ang. Nei & Li similarity): $\frac{d_{11}}{(d_{11}+d_{01})+(d_{11}+d_{10})}$
18. Współczynnik podobieństwa Sokala & Sneatha (I) (ang. Sokal & Sneath similarity (I)): $\frac{d_{11}}{d_{11}+2d_{01}+2d_{10}}$
19. Współczynnik podobieństwa Sokala & Sneatha (II) (ang. Sokal & Sneath similarity (II)): $\frac{2(d_{11}+d_{00})}{2d_{11}+d_{01}+d_{10}+2d_{00}}$
20. Współczynnik podobieństwa Sokala & Sneatha (III) (ang. Sokal & Sneath similarity (III)): $\frac{d_{11}+d_{00}}{d_{01}+d_{10}}$
21. Współczynnik podobieństwa Sokala & Sneatha (IV) (ang. Sokal & Sneath similarity (IV)): $\frac{\frac{d_{11}}{(d_{11}+d_{01})}+\frac{d_{11}}{(d_{11}+d_{10})}+\frac{d_{00}}{(d_{01}+d_{00})}+\frac{d_{00}}{(d_{10}+d_{00})}}{4}$
22. Współczynnik podobieństwa Sokala & Sneatha (V) (ang. Sokal & Sneath similarity (V)):$\frac{d_{11}d_{00}}{((d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00}))^{0,5}}$
23. Współczynnik podobieństwa Sokala & Michenera (ang. Sokal & Michener similarity): $\frac{d_{11}+d_{00}}{d_{11}+d_{01}+d_{10}+d_{00}}$
24. Współczynnik podobieństwa Rogera & Tanimoto (ang. Roger & Tanimoto similarity): $\frac{d_{11}+d_{00}}{d_{11}+2(d_{01}+d_{10})+d_{00}}$
25. Współczynnik podobieństwa Gowera & Legendre'a (ang. Gower & Legendre similarity): $\frac{d_{11}+d_{00}}{d_{11}+0,5(d_{01}+d_{10})+d_{00}}$
26. Współczynnik podobieństwa Russera & Rao (ang. Russer & Rao similarity): $\frac{d_{11}}{d_{11}+d_{01}+d_{10}+d_{00}}$
27. Współczynnik podobieństwa cosinusów (ang. Cosine similarity): $\frac{d_{11}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})}^2}$
28. Współczynnik podobieństwa Gilberta & Wells'a (ang. Gilbert & Wells similarity): $\log d_{11} - \log n - \log\left(\frac{d_{11}+d_{01}}{n}\right) - \log\left(\frac{d_{11}+d_{10}}{n}\right)$
29. Współczynnik podobieństwa Ochiai (I) (ang. Ochiai similarity (I)): $\frac{d_{11}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})}}$

30. Współczynnik podobieństwa Ochiai (II) (ang. Ochiai similarity (II)): $\frac{d_{11}d_{00}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}}$
31. Współczynnik podobieństwa Forbesi (ang. Forbesi similarity):$\frac{nd_{11}}{(d_{11}+d_{01})(d_{11}+d_{10})}$
32. Współczynnik podobieństwa Fossum (ang. Fossum similarity):$\frac{n(d_{11}-0{,}5)^2}{(d_{11}+d_{01})(d_{11}+d_{10})}$
33. Współczynnik podobieństwa Sorgenfrei (ang. Sorgenfrei similarity): $\frac{{d_{11}}^2}{(d_{11}+d_{01})(d_{11}+d_{10})}$
34. Współczynnik podobieństwa Mountforda (ang. Mountford similarity): $\frac{d_{11}}{0{,}5(d_{11}d_{01}+d_{11}d_{10})+d_{01}d_{10}}$
35. Współczynnik podobieństwa Otsuki (ang. Otsuka similarity):$\frac{d_{11}}{((d_{11}+d_{01})(d_{11}+d_{10}))^{0,5}}$
36. Współczynnik podobieństwa McConnaughey'a (ang. McConnaughey similarity): $\frac{{d_{11}}^2-d_{01}d_{10}}{(d_{11}+d_{01})(d_{11}+d_{10})}$
37. Współczynnik podobieństwa Tarwida (ang. Tarwid similarity): $\frac{nd_{11}-(d_{11}+d_{01})(d_{11}+d_{10})}{nd_{11}+(d_{11}+d_{01})(d_{11}+d_{10})}$
38. Współczynnik podobieństwa Kulczyńskiego (I) (ang. Kulczynski similarity (I)): $\frac{d_{11}}{d_{01}+d_{10}}$
39. Współczynnik podobieństwa Kulczyńskiego (II) (ang. Kulczynski similarity (II)): $\frac{\frac{d_{11}}{2}(2d_{11}+d_{01}+d_{10})}{(d_{11}+d_{01})(d_{11}+d_{10})}$
40. Współczynnik podobieństwa Johnsona (ang. Johnson similarity): $\frac{d_{11}}{d_{11}+d_{01}}+\frac{d_{11}}{d_{11}+d_{10}}$
41. Współczynnik podobieństwa Dennisa (ang. Dennis similarity): $\frac{d_{11}d_{00}-d_{01}d_{10}}{\sqrt{n(d_{11}+d_{01})(d_{11}+d_{10})}}$
42. Współczynnik podobieństwa Simpsona (ang. Simpson similarity): $\frac{d_{11}}{\min(d_{11}+d_{01},d_{11}+d_{10})}$
43. Współczynnik podobieństwa Brauna & Banqueta (ang. Braun & Banquet similarity): $\frac{d_{11}}{\max(d_{11}+d_{01},d_{11}+d_{10})}$
44. Współczynnik podobieństwa Fagera & McGowana (ang. Fager & McGowan similarity): $\frac{d_{11}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})}}-\frac{\max(d_{11}+d_{01},d_{11}+d_{10})}{2}$
45. Współczynnik podobieństwa Forbesa (ang. Forbes similarity): $\frac{nd_{11}-(d_{11}+d_{01})(d_{11}+d_{10})}{n\min(d_{11}+d_{01},d_{11}+d_{10})-(d_{11}+d_{01})(d_{11}+d_{10})}$
46. Współczynnik podobieństwa Gowera (ang. Gower similarity): $\frac{d_{11}+d_{00}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}}$

47. Współczynnik podobieństwa Pearsona (I) (ang. Pearson similarity (I)): χ^2,

    gdzie $\chi^2 = \frac{n(d_{11}d_{00}-d_{01}d_{10})^2}{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}$

48. Współczynnik podobieństwa Pearsona (II) (ang. Pearson similarity (II)): $(\frac{\chi^2}{n+\chi^2})^{1/2}$

49. Współczynnik podobieństwa Pearsona (III) (ang. Pearson similarity (III)): $(\frac{\rho}{n+\rho})^{1/2}$,

    gdzie $\rho = \frac{d_{11}d_{00}-d_{01}d_{10}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}}$

50. Współczynnik podobieństwa Pearsona & Herona (I) (ang. Pearson & Heron similarity (I)): $\frac{d_{11}d_{00}-d_{01}d_{10}}{\sqrt{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}}$

51. Współczynnik podobieństwa Pearsona & Herona (II) (ang. Pearson & Heron similarity (II)): $\cos(\frac{\pi\sqrt{d_{01}+d_{10}}}{\sqrt{d_{11}+d_{00}}+\sqrt{d_{01}+d_{10}}})$

52. Współczynnik podobieństwa Cole'a (ang. Cole similarity): $\frac{\sqrt{2}(d_{11}d_{00}-d_{01}d_{10})}{\sqrt{(d_{11}d_{00}-d_{01}+d_{10})^2-(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}}$

53. Współczynnik podobieństwa Stilesa (ang. Stiles similarity): $\log_{10}\frac{n(|d_{11}d_{00}-d_{01}d_{10}|-\frac{n}{2})^2}{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}$

54. Współczynnik podobieństwa Yule'a (I) (ang. Yule similarity (I)): $\frac{d_{11}d_{00}-d_{01}d_{10}}{d_{11}d_{00}+d_{01}d_{10}}$

55. Współczynnik podobieństwa Yule'a (II) (ang. Yule similarity (II)): $\frac{\sqrt{d_{11}d_{00}}-\sqrt{d_{01}d_{10}}}{\sqrt{d_{11}d_{00}}-\sqrt{d_{01}d_{10}}}$

56. Współczynnik podobieństwa Tanimoto'a (ang. Tanimoto similarity): $\frac{d_{11}}{(d_{11}+d_{01})+(d_{11}+d_{10})-d_{11}}$

57. Współczynnik podobieństwa Dispersona (ang. Disperson similarity): $\frac{d_{11}d_{00}-d_{01}d_{10}}{(d_{11}+d_{01}+d_{10}+d_{00})^2}$

58. Współczynnik podobieństwa Hamanna (ang. Hamann similarity): $\frac{(d_{11}+d_{00})-(d_{01}+d_{10})}{d_{11}+d_{01}+d_{10}+d_{00}}$

59. Współczynnik podobieństwa Michaela (ang. Michael similarity): $\frac{4(d_{11}d_{00}-d_{01}d_{10})}{(d_{11}+d_{00})^2-(d_{01}+d_{10})^2}$

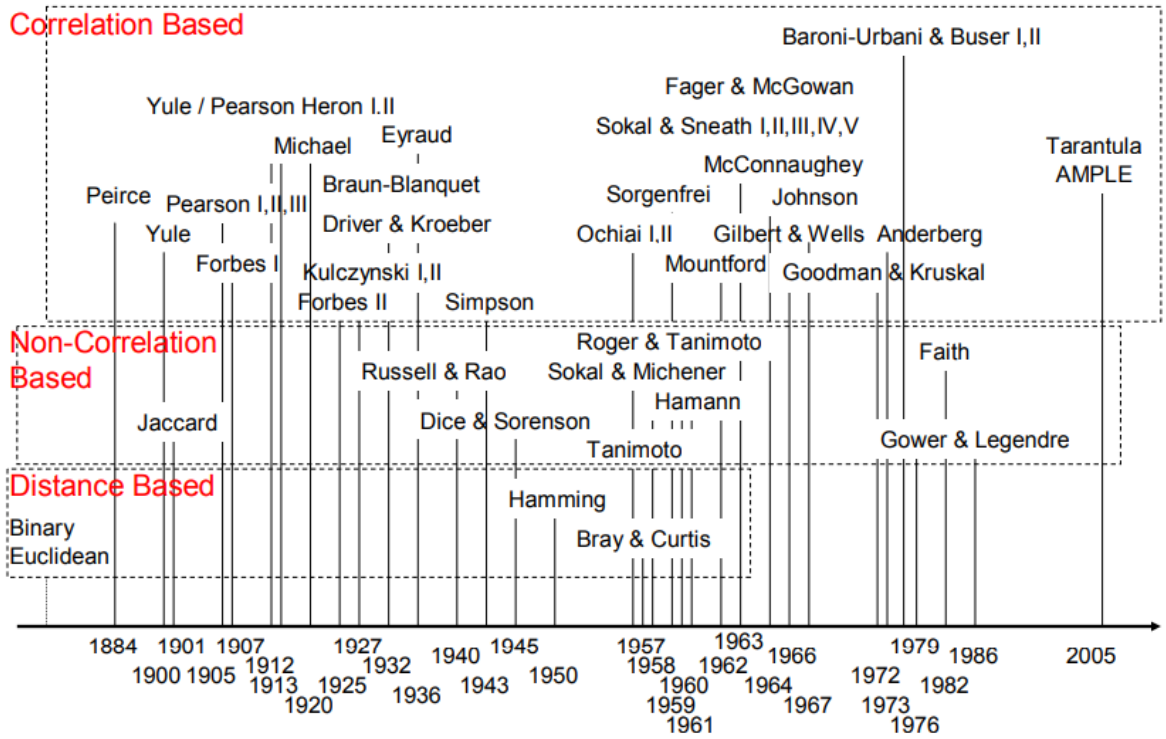
60. Współczynnik podobieństwa Goodmana & Kruskala (ang. Goodman & Kruskal similarity): $\frac{\sigma-\sigma'}{2n-\sigma'}$, gdzie $\sigma = \max(d_{11}, d_{01}) + \max(d_{10}, d_{00}) + \max(d_{11}, d_{10}) + \max(d_{01}, d_{00})$,
    $\sigma' = \max(d_{11} + d_{10},\quad d_{01} + d_{00}) + \max(d_{11} + d_{01},\quad d_{10} + d_{00})$

61. Współczynnik podobieństwa Anderberga (ang. Anderberg similarity): $\frac{\sigma-\sigma'}{2n}$

62. Współczynnik podobieństwa Baroni – Urbani & Busera (I) (ang. Baroni – Urbani & Buser similarity (I)): $\frac{\sqrt{d_{11}d_{00}}+d_{11}}{\sqrt{d_{11}d_{00}}+d_{11}+d_{01}+d_{10}}$

63. Współczynnik podobieństwa Baroni – Urbani & Busera (II) (ang. Baroni – Urbani & Buser similarity (II)): $\frac{\sqrt{d_{11}d_{00}}+d_{11}-(d_{01}+d_{10})}{\sqrt{d_{11}d_{00}}+d_{11}+d_{01}+d_{10}}$
64. Współczynnik podobieństwa Peirce'a (ang. Peirce similarity): $\frac{d_{11}d_{01}+d_{01}d_{10}}{d_{11}d_{01}+2d_{01}d_{10}+d_{10}d_{00}}$
65. Współczynnik podobieństwa Eyraud'a (ang. Eyraud similarity): $\frac{n^2(nd_{11}-(d_{11}+d_{01})(d_{11}+d_{10}))}{(d_{11}+d_{01})(d_{11}+d_{10})(d_{01}+d_{00})(d_{10}+d_{00})}$

Poniższa ilustracja przedstawia wymienione współczynniki i odległości wraz z oznaczeniem czasu ich powstania.



Rysunek 1. Oś czasu powstawania kolejnych binarnych miar podobieństwa. [2]

Znane są także współczynniki różnic (niepodobieństwa) (ang. dissimilarity measures) [22] [23]:

1. Współczynnik różnicy Dice'a – Sørensena (ang. Dice – Sørensen dissimilarity): $\frac{d_{01}+d_{10}}{2d_{11}+d_{01}+d_{10}}$
2. Współczynnik różnicy Jaccarta – Needhama (ang. Jaccard – Needham dissimilarity): $\frac{d_{01}+d_{10}}{d_{11}+d_{01}+d_{10}}$
3. Współczynnik różnicy Kulsińskiego (ang. Kulsinski dissimilarity): $\frac{d_{01}+d_{10}-d_{11}+n}{d_{01}+d_{10}+n}$

4. Współczynnik różnicy Rogera – Tanimoto (ang. Rogers – Tanimoto dissimilarity): $\frac{2(d_{01}+d_{10})}{d_{11}+2(d_{01}+d_{10})+d_{00}}$
5. Współczynnik różnicy Russera – Rao (ang. Russer – Rao dissimilarity): $\frac{n-d_{11}}{n}$
6. Współczynnik różnicy Sokala – Sneatha (ang. Sokal – Sneath dissimilarity): $\frac{2(d_{01}+d_{10})}{d_{11}+2(d_{01}+d_{10})}$
7. Współczynnik różnicy Yule'a (ang. Yule dissimilarity): $\frac{d_{01}d_{10}}{d_{11}d_{00}+d_{01}d_{10}}$

Definicja 2.2. [28] Analiza skupień

Analiza skupień jest procesem przekształcającym zbiór na podzbiory zwane „skupieniami". Własnością skupień jest fakt, że dane wewnątrz podzbioru różnią się od siebie mniej niż dane należące do innych podzbiorów. Można więc powiedzieć, że proces tej analizy prowadzi do rozpoznania pewnych struktur w rozważanym zbiorze. Procesem prowadzącym do podziału zbioru na skupienia jest odpowiednio dobrana miara podobieństwa bądź odmienności. Wśród wielu zastosowań tego podejścia, warto wyróżnić wykorzystanie analizy skupień do redukcji dużych zbiorów danych, tzw. big data – umożliwia bowiem zastąpienie zbioru jego reprezentatywną próbką. Formalizując, jeśli przyjmiemy, że dany jest zbiór n obiektów $X = \{x_1, x_2, \dots, x_n\}$, przy czym każdy obiekt jest opisany m – wymiarowym wektorem $x_i = (x_{i1} \dots x_{im})^T$, gdzie x_{ij} oznacza wartość j-tej cechy w obiekcie x_i. Celem analizy skupień jest podział zbioru X na $k < n$ grup $C = \{C_1, \dots, C_k\}$, gdzie i-ta grupa C_i nazywana jest skupieniem (klastrem). Podział ten spełnia trzy warunki:

i. Każde skupienie zawiera przynajmniej jeden obiekt, $C_j \neq \emptyset, j = 1, \dots, k$;
ii. Każdy obiekt należy do pewnego skupienia, $\bigcup_{j=1}^{k} C_j = X$
iii. Każdy obiekt należy do dokładnie jednego skupienia, $C_{j_1} \cap C_{j_2} = \emptyset, j_1 \neq j_2$.

Analiza skupień jest jedną z metod uczenia nienadzorowanego.

Definicja 2.3. [28] Hierarchiczne metody analizy skupień

Metody hierarchiczne oparte są na stopniowym łączeniu bądź rozdzielaniu obiektów. Wynikiem tego procesu reprezentującym jego efekty jest struktura drzewiasta nazywana dendrogramem. Metody hierarchiczne można podzielić na techniki rozdrobnieniowe i aglomeracyjne. W metodach rozdrobnieniowych analizę rozpoczyna się od jednego skupienia złożonego z całego zbioru. Podczas kolejnych etapów procesu zbiór jest dzielony na mniejsze – zgodnie z rosnącym podobieństwem wewnątrz poszczególnych grup. Metody aglomeracyjne rozpoczynają od n zbiorów obserwacji – każdy obiekt zbioru uznany jest jako osobne skupienie. Klastry łączone są ze sobą wraz ze wzrastającym stopniem odmienności pomiędzy obiektami. Łączenie bądź dzielenie skupień

odbywa się pod wpływem różnych algorytmów, nie zależnie od wybranej techniki. Wyróżnia się m.in.

a) Metodę pojedynczego wiązania lub najbliższego sąsiedztwa (ang. Single linkage): Odległością pomiędzy dwoma klastrami jest odległość między dwoma najbliższymi obiektami należącymi do różnych skupień. Do znalezienia optymalnego rozwiązania tego zadania stosuje się algorytm oparty na minimalnym drzewie rozpinającym.
b) Metoda pełnego wiązania lub najdalszego sąsiada (ang. Complete linkage): Odległością pomiędzy dwoma klastrami jest odległość między dwoma najdalszymi obiektami należącymi do różnych skupień. Metoda ta jest odpowiednia do zbiorów, w których obiekty tworzą dobrze izolowane i zwarte skupienia.
c) Metoda środkowego wiązania lub środkowego sąsiada (ang. Median linkage): Odległością pomiędzy dwoma klastrami jest odległość między dwiema medianami obiektów należącymi do różnych skupień.

Definicja 2.4. [28] Relacyjne metody analizy skupień

Podejście to nie bazuje na reprezentacji wektorowej obiektów, a na opisaniu ich za pomocą relacji podobieństwa bądź odmienności między parami obiektów, czego wynikiem jest wytworzenie macierzy podobieństwa wewnątrz zbioru.

Definicja 2.5. [24][25][28] Propagacja powinowactwa

Jest to algorytm z grupy algorytmów relacyjnych. Jako dane wejściowe wprowadzane są wyniki obliczonych podobieństw między parami wyrazów, czyli macierz podobieństwa składająca się z wyrazów s_{ij} . Jako dane wyjściowe zwrócona zostaje lista klastrów – skupień wraz z wyznaczonymi dla nich etykietami, nazywanymi też archetypami. Algorytm propagacji powinowactwa opiera się na maksymalizacji funkcji:

$$E(c) = -\sum_{i=1}^{m} s_{i,c_i} + \sum_{j=1}^{m} \delta_j(c),$$

gdzie $c = (c_1, \dots, c_m)$ jest zbiorem etykiet – etykieta c_i wskazuje reprezentanta wyrazu i, natomiast

$$\delta_j(c) = \begin{cases} -\infty, jeżeli\ c_j \neq j\ oraz\ \exists i: c_i = j \\ 0\ w\ p.p. \end{cases}$$

jest kosztem przypisywanym w sytuacji, gdy wyraz i wybiera wyraz $j = c_i$ za swój prototyp, natomiast wyraz j nie uznaje siebie za prototyp, więc $c_j \neq j$. Okazuje się więc, że warunek, jaki musi być spełniony przez prototyp, to warunek $c_j = j$, co oznacza, że zarówno kilka wyrazów wybiera wyraz j jako swojego reprezentanta, jak i wyraz j jest prototypem. Rozwiązaniem problemu

maksymalizacji powyższej funkcji jest więc wymiana informacji pomiędzy wyrazami. Informacja r_{ij} wysyłana od wyrazu i do j jest równoznaczna z byciem prototypem (responsibility) dla wyrazu i. Z kolei informacja a_{ij} wysyłana od wyrazu j do i jest równoznaczna ze zgłoszeniem gotowości do bycia prototypem dla wyrazu i (availability). Główną zasadą działania algorytmu jest więc ciągła iteracyjna aktualizacja wartości odpowiedzialności i gotowości (wymiana komunikatów), która maksymalizuje sumę podobieństw wyrazów do przydzielonych im etykiet. Jako warunek stopu przyjmujemy ustaloną ilość iteracji algorytmu bądź osiągnięcie minimalnej dokonanej poprawy dokładności wyniku. Ważną cechą algorytmu propagacji powinowactwa jest to, że działa on w sposób nienadzorowany, co oznacza, że nie definiujemy tu ilości grup do rozpoznania – algorytm wyznacza tę ilość automatycznie.

Definicja 2.6. [28] Kombinatoryczne metody analizy skupień

Metody kombinatoryczne skupiają się na problemie wyboru właściwego podziału zbioru. Ten problem należy do kategorii NP-zupełnych i jest zadaniem optymalizacji kombinatorycznej. Do ewaluacji dopuszczalnych podziałów wykorzystuje się kryteria grupowania i dopasowane do nich metody ich optymalizacji. Przykładami takich kryteriów są te oparte za odmienności – homogeniczności i separacji, inne oparte na minimalizacji śladu macierzy kowariancji wewnątrzgrupowych bądź na aproksymacji macierzy danych.

Definicja 2.7. [29] Algorytm K-średnich

Jedna z kombinatorycznych metod analizy skupień. Metoda ta wykorzystuje pojęcie „centroidu skupiska", który jest wyznaczany jako średnia lokalizacji obiektów wewnątrz wyznaczonego skupienia. Jest to metoda, którą stosuje się w przypadku gdy jawna jest ilość klastrów. Na potrzeby definicji założono, że ilość klastrów wynosi k. Algorytm przebiega w następujący sposób:

i. Należy wyznaczyć startowe k centroidów – można zrobić to wybierając k losowych punktów, wybór może zostać dokonany również wybierając k losowych punktów należących do rozważanego zbioru.
ii. Dla każdego z obiektów należy wyznaczyć ich odległość od każdego z k centroidów oraz dopasować do centroidu z którym łączy go najmniejsza odległość.
iii. Dla skupień powstałych w poprzednim punkcie wyznaczamy nowe położenia centroidów.
iv. Kroki ii. oraz iii. powtarzamy do momentu ustabilizowania się skupień.

Definicja 2.8. [30] Analiza syntaktyczna

Analiza syntaktyczna polega na podziale rozważanych obiektów na obiekty prostsze. Proces ten przeprowadza się do momentu otrzymania tzw. składowych pierwotnych, czyli części

niepodzielnych, takich dla których jest możliwe jednoznaczne ich rozpoznanie i zdefiniowanie. Przed rozpoczęciem analizy syntaktycznej należy wyodrębnić składowe pierwotne oraz relacje, jakie zachodzą między nimi. Proces ten nazywany jest rozpoznaniem strukturalnym. Sposób ten ma szerokie zastosowanie w analizie i rozpoznawaniu obrazów. Metodami rozpoznawania strukturalnego dedykowanymi do rozpoznawania obrazów są metody ciągowe, metody grafowe, metody drzewowe oraz analiza scen i faktur. Za pomocą tych metod można zdefiniować mechanizm generujący reprezentację rozważanych obiektów. Ten mechanizm nazywany jest gramatyką, a zbiór wszystkich reprezentacji obiektów przez nią generowanych nazywany jest językiem.

Definicja 2.9. [30] Metody ciągowe

Jedna z metod rozpoznawania strukturalnego. W tej metodzie obraz reprezentowany jest jako ciąg. Jest to równoznaczne z tym, że wyróżniamy jeden rodzaj relacji, jaki zachodzi pomiędzy składowymi pierwotnymi – konkatenację, czyli dołączanie kolejnych elementów. Tak scharakteryzowane ciągi nazywamy językami ciągowymi. Przykładami języków ciągowych są języki oparte na kodach łańcuchowych (Freemana), języki opisu obrazów (Shawa) oraz języki opisu cech kształtów (Jakubowskiego).

Definicja 2.10. [30] Kody łańcuchowe Freemana

Jedna z syntaktycznych metod rozpoznawania obrazów. Jej cechą charakterystyczną jest prostota reprezentacji obiektów oraz liniowa złożoność procedury rozpoznającej. Zasadą tworzenia kodów Freemana jest podział obiektów na składowe i zapisanie ich za pomocą wektorów. Poniższa ilustracja prezentuje a) składowe pierwotne kodu Freemana oraz b) zbudowane za ich pomocą kontury wybranych liter.

Rys. 2. Wizualizacja zasady tworzenia kodów Freemana. [30]

Rozpoznania tak zdefiniowanych obrazów zwykle dokonuje się za pomocą automatu opartego na funkcji przejścia dla ciągu symboli.

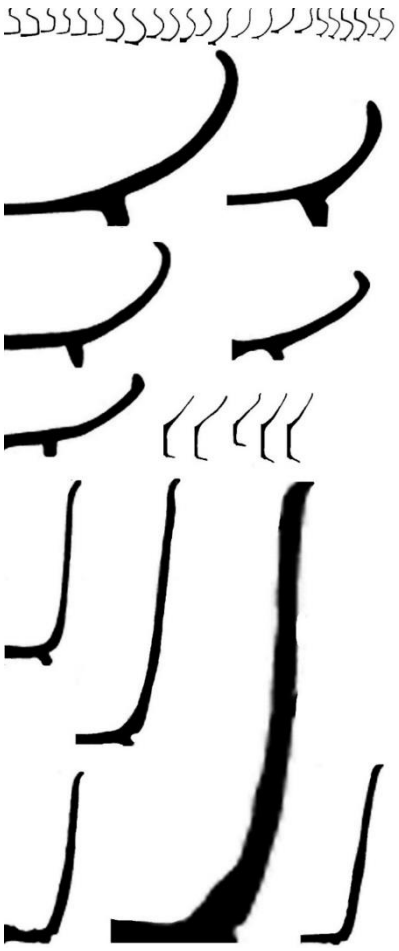
Rozdział 3.
Opis i wstępna obróbka zbioru testowego

W tym rozdziale przedstawiona zostanie charakterystyka zbioru reprezentującego problem. Znajduje się tu także opis obróbki wstępnej zbioru.

Zbiór testowy składa się z 38 krzywych, stworzonych metodą ręczną i reprezentujących profile ceramiczne, ponumerowanych od 0.jpg do 37.jpg. W szczególności, są to:

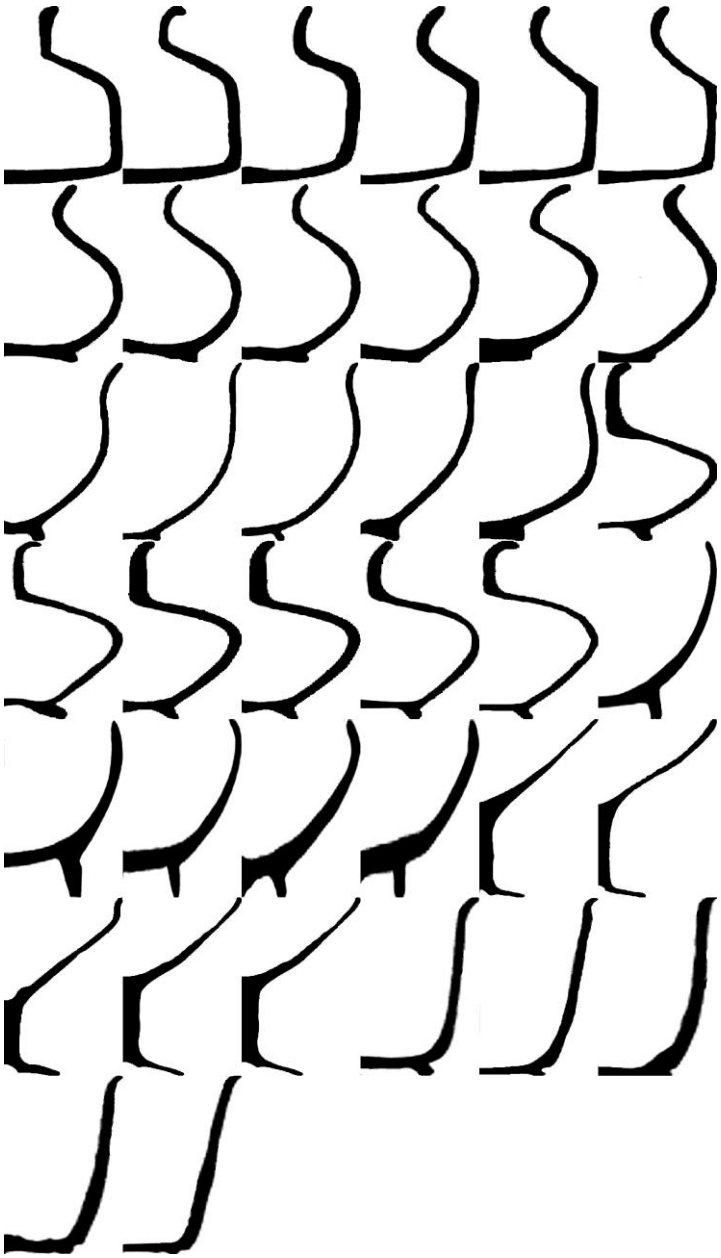
- 0-5: alabastry,
- 6-11: amphoriskosy,
- 12-16: misy,
- 17-22: lekyty,
- 23-27: wazy,
- 28-32: kielichy,
- 33-37: naczynia pionowe duże.

Poniższa ilustracja przedstawia te krzywe zachowując skalę rozmiaru.



Rysunek 3. Zbiór testowy, profile w oryginalnych proporcjach.

Przed rozpoczęciem analizy należy przygotować odpowiednio dane wejściowe poprzez ich normalizację do stałego wymiaru 300x200 pikseli. Preferowanym i wystarczającym formatem obrazów reprezentujących krzywe jest format .jpg. Poniżej znajduje się ilustracja przedstawiająca znormalizowane profile.



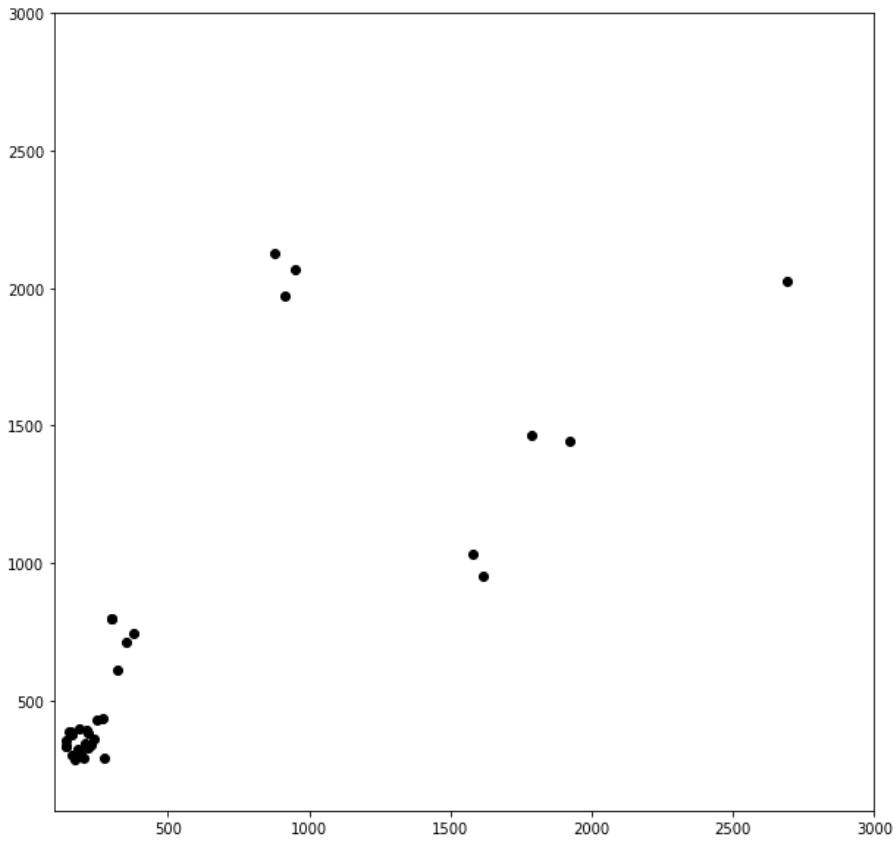
Rysunek 4. Zbiór testowy, profile po normalizacji w rozmiarach 300x200 pikseli

Kolejnym etapem obróbki wstępnej danych wejściowych jest konwersja obrazu na macierz pikseli. Ponieważ obraz składa się jedynie z białych i czarnych pikseli, otrzymana macierz jest macierzą binarną. Tak przygotowany zbiór gotowy jest do przeprowadzenia analizy oraz procesu klastrowania. Na potrzeby niektórych z testowanych metod macierz przekształcana jest w procesie konkatenacji wierszy na wektor.

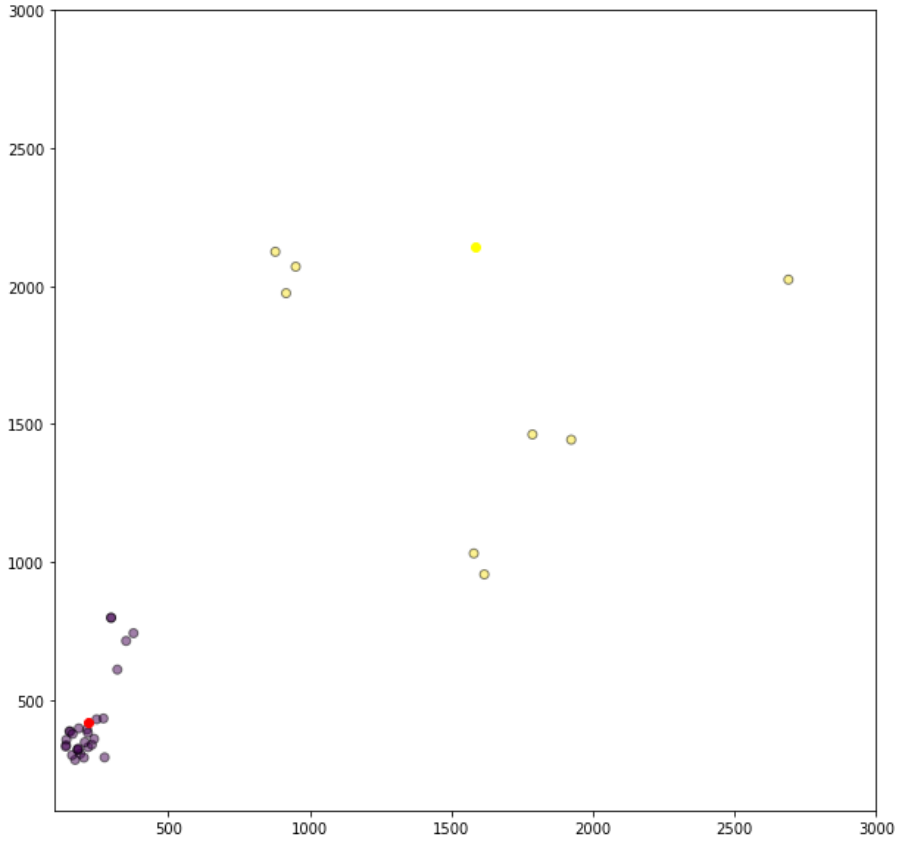
Rozdział 4.
Analiza skupień

W tym rozdziale znajduje się zastosowanie analizy skupień w praktyce oraz jego ewaluacja pod kątem prawidłowości wyników. Znajduje się tu także ocena zastosowania tego podejścia do zbioru profili ceramicznych.

Pierwszym etapem, tzw. „wstępnym" w autorskim podejściu do analizy skupień za pomocą miar podobieństwa jest analiza rozmiarów oryginalnych danych wejściowych. Jest ona istotna, ponieważ dzięki temu nie traci się informacji o oryginalnym rozmiarze. Przyjęto zasadę, że jeśli chociaż jeden z profili zostanie rozpoznany jako 3 razy większy niż inne to, bez dalszej analizy rozmiarów, następuje rozdzielenie grupy profili na dwa klastry, podział następuje po wielkości. Do jego wykonania zdecydowano się wybrać metodę analizy skupień K-średnich (ang. KMeans). Podjęto taką decyzję, ponieważ w tym przypadku już na początku znana jest liczba oczekiwanych klastrów – 2. Nie ma więc wątpliwości, w którym momencie przyjąć punkt odcięcia w procesie klastrowania metodą KMeans. Na poniższych ilustracjach zobaczyć można rozmieszczenie punktów reprezentujących rozmiary poszczególnych profili oraz ich podział na klastry pod kątem rozmiaru.



Rysunek 5. Rozmieszczenie profili wejściowych pod względem ich oryginalnego rozmiaru.



Rysunek 6. Wizualizacja podziału profili na dwie grupy względem rozmiaru. Intensywnymi kolorami zaznaczone zostały centroidy.

Wynik analizy - poprawny:

```
{0: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15,
16,17, 18, 19, 20, 21, 22, 28, 29, 30, 31, 32], dtype=int64),
 1: array([23, 24, 25, 26, 27, 33, 34, 35, 36, 37], dtype=int64)}
```

Po obróbce danych oraz ich wstępnej analizie otrzymujemy dwa zbiory profili, które od tego momentu będą analizowane niezależnie od siebie. Również od tego momentu prace przebiegać będą na macierzach binarnych reprezentujących profile. Z tego też powodu efektownymi metodami pracy na danych są te dedykowane do danych binarnych.

Jako miary podobieństwa, zastosowane zostaną:

1. Odległość Hamminga.
2. Odległość Vari'ego.
3. Współczynnik różnicy Rogera – Tanimoto.
4. Odległość różnicy rozmiaru.
5. Odległość różnicy wzorca.
6. Współczynnik różnicy Jaccarta – Needhama.

7. Współczynnik różnicy Sokala – Sneatha.
8. Współczynnik różnicy Russera – Rao.
9. Współczynnik podobieństwa Sokala – Sneatha (IV).
10. Współczynnik podobieństwa Sokala – Sneatha (V).
11. Współczynnik różnicy Yule'a.
12. Odległość Hellingera.
13. Współczynnik podobieństwa Gowera.
14. Współczynnik podobieństwa Goodmana & Kruskala.

Wybierając miary do analizy, skupiono się na ich różnorodności. Odległość Hamminga skupia się tylko na informacji o różnicy obiektów. Odległość Vari'ego i współczynnik różnicy Rogera – Tanimoto biorą pod uwagę pełne spectrum informacji o podobieństwach i różnicach obiektów. Odległość różnicy rozmiaru i różnicy wzorca to miary badające kwadraty informacji. Współczynniki różnicy Jaccarta – Needhama oraz Sokala – Sneatha nie biorą pod uwagę informacji o ilości nałożonych na siebie białych pikseli na obu obrazkach, korzystają więc z niepełnej informacji. Współczynnik różnicy Russera – Rao korzysta tylko jednej części informacji – ilości czarnych pikseli nałożonych na siebie w obu obiektach. Współczynnik podobieństwa Sokala – Sneatha (IV) opiera się na proporcjach, a współczynniki podobieństwa Sokala – Sneatha (V) oraz Gowera biorą pod uwagę pierwiastek z informacji. Współczynnik różnicy Yule'a opiera się na iloczynach informacji, a współczynnik podobieństwa Goodmana & Kruskala skupia się na analizie maksimów wśród informacji.

Całość eksperymentu jest wykonana poprzez stworzenie kodu w języku programowania Python, wersja 3.6, kod uruchamiany był przy użyciu aplikacji/interpretera Jupyter Notebook.

Poniżej znajduje się wynik obliczonych miar podobieństwa przed przystąpieniem do procesu propagacji powinowactwa. Widać, że między wskaźnikami występują różnice.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
0	0	0,168358	0,318692	0,314181	0,246101	0,263336	0,331062	0,299681	0,355601	0,385165	0,419264	0,430018	0,452412	0,425383	0,422091	0,471552	0,436075	0,518299	0,440323	0,478084	0,481147	0,446883	0,403746	0,487064	0,472972	0,478451	0,497427	0,486645
1	0,168358	0	0,234827	0,259533	0,199388	0,215587	0,293588	0,266297	0,296374	0,343868	0,362602	0,408214	0,44626	0,424866	0,420866	0,464766	0,423852	0,517072	0,471415	0,501442	0,505418	0,479954	0,44859	0,494939	0,481147	0,486645	0,503218	0,513218
2	0,318692	0,234827	0	0,183112	0,176569	0,188105	0,215481	0,28277	0,216939	0,253377	0,315056	0,360409	0,451433	0,4218	0,415559	0,464963	0,401471	0,504654	0,479781	0,504952	0,509428	0,47268	0,459425	0,49863	0,485976	0,488969	0,507514	0,509428
3	0,314181	0,259533	0,183112	0	0,18553	0,187859	0,270993	0,332569	0,267823	0,27027	0,318715	0,323507	0,461262	0,418201	0,420886	0,478335	0,425073	0,523549	0,485441	0,510168	0,515495	0,488246	0,457901	0,49197	0,478432	0,489483	0,511	0,515495
4	0,246101	0,199388	0,176569	0,18553	0	0,128355	0,221353	0,285616	0,24181	0,301053	0,324467	0,395529	0,448911	0,428208	0,420367	0,46755	0,418285	0,51314	0,463075	0,490129	0,494845	0,468235	0,438984	0,487083	0,473418	0,48324	0,5003	0,505418
5	0,263336	0,215587	0,188105	0,187859	0,128355	0	0,240779	0,310311	0,225193	0,273754	0,313612	0,366191	0,413831	0,390925	0,381179	0,432704	0,38857	0,492349	0,447185	0,474331	0,480031	0,45679	0,418514	0,447165	0,432315	0,439796	0,461992	0,464472
6	0,331062	0,293588	0,215481	0,270993	0,221353	0,240779	0	0,17321	0,182121	0,247791	0,274573	0,358278	0,417972	0,413894	0,391162	0,436013	0,386056	0,439207	0,40758	0,43876	0,440526	0,384143	0,385621	0,470519	0,455519	0,460058	0,466826	0,439796
7	0,299681	0,266297	0,28277	0,332569	0,285616	0,310311	0,17321	0	0,240727	0,301919	0,306783	0,384643	0,40872	0,416803	0,393123	0,427879	0,386164	0,42348	0,400682	0,433686	0,438882	0,378663	0,384687	0,464472	0,446823	0,453489	0,450293	0,464472
8	0,355601	0,296374	0,216939	0,267823	0,24181	0,225193	0,182121	0,240727	0	0,174072	0,264818	0,27549	0,394026	0,371822	0,353706	0,414502	0,34261	0,418201	0,406882	0,432868	0,434728	0,384861	0,380414	0,438984	0,421945	0,427879	0,447808	0,44626
9	0,385165	0,343868	0,253377	0,27027	0,301053	0,273754	0,247791	0,301919	0,174072	0	0,286865	0,213887	0,375107	0,36539	0,344805	0,399018	0,32442	0,417325	0,403236	0,430655	0,436319	0,39032	0,390536	0,41891	0,400363	0,406459	0,426395	0,412595
10	0,419264	0,362602	0,315056	0,318715	0,324467	0,313612	0,274573	0,306783	0,264818	0,286865	0	0,282869	0,380742	0,330295	0,303647	0,394799	0,272536	0,424411	0,367636	0,38131	0,389888	0,401023	0,349835	0,441194	0,428291	0,436095	0,453349	0,461992
11	0,430018	0,408214	0,360409	0,323507	0,395529	0,366191	0,358278	0,384643	0,27549	0,213887	0,282869	0	0,409754	0,345535	0,331201	0,436462	0,314654	0,432438	0,408551	0,426807	0,431229	0,410891	0,380916	0,447205	0,431332	0,439714	0,468978	0,453885
12	0,452412	0,44626	0,451433	0,461262	0,448911	0,413831	0,417972	0,40872	0,394026	0,375107	0,380742	0,409754	0	0,277125	0,304175	0,154669	0,298838	0,458297	0,417701	0,446039	0,451113	0,468763	0,427013	0,366458	0,335322	0,354113	0,375239	0,366458
13	0,425383	0,424866	0,4218	0,418201	0,428208	0,390925	0,413894	0,416803	0,371822	0,36539	0,330295	0,345535	0,277125	0	0,142731	0,248329	0,19882	0,437113	0,361059	0,383991	0,393746	0,445576	0,366925	0,37932	0,353074	0,364008	0,402216	0,418285
14	0,422091	0,420866	0,415559	0,420886	0,420367	0,381179	0,391162	0,393123	0,353706	0,344805	0,303647	0,331201	0,304175	0,142731	0	0,285837	0,156682	0,41047	0,328295	0,348927	0,359244	0,412614	0,328877	0,375305	0,347335	0,361394	0,389543	0,375305
15	0,471552	0,464766	0,464963	0,478335	0,46755	0,432704	0,436013	0,427879	0,414502	0,399018	0,394799	0,436462	0,154669	0,248329	0,285837	0	0,286131	0,478258	0,439572	0,466081	0,47196	0,489787	0,450433	0,386641	0,361618	0,36441	0,399872	0,386641
16	0,436075	0,423852	0,401471	0,425073	0,418285	0,38857	0,386056	0,386164	0,34261	0,32442	0,275356	0,314654	0,298838	0,19882	0,156682	0,286131	0	0,397479	0,343159	0,357964	0,365011	0,405358	0,346014	0,403321	0,379561	0,385643	0,421032	0,403321
17	0,518299	0,517072	0,504654	0,523549	0,51314	0,492349	0,439207	0,42348	0,418201	0,417325	0,424411	0,432438	0,458297	0,437113	0,41047	0,478258	0,397479	0	0,308501	0,262833	0,274201	0,289087	0,319657	0,489882	0,480859	0,487732	0,486091	0,475009
18	0,440323	0,471415	0,479781	0,485441	0,463075	0,447185	0,40758	0,400682	0,406882	0,403236	0,367636	0,408551	0,417701	0,361059	0,328295	0,439572	0,343159	0,308501	0	0,165671	0,174239	0,222196	0,202759	0,475009	0,457841	0,458851	0,459168	0,445576
19	0,478084	0,501442	0,504952	0,510168	0,490129	0,474331	0,43876	0,433686	0,432868	0,430655	0,38131	0,426807	0,446039	0,383991	0,348927	0,466081	0,357964	0,262833	0,165671	0	0,128363	0,23035	0,237627	0,492349	0,481243	0,479492	0,486435	0,489787
20	0,481147	0,505418	0,509428	0,515495	0,494845	0,480031	0,440526	0,438882	0,434728	0,436319	0,389888	0,431229	0,451113	0,393746	0,359244	0,47196	0,365011	0,274201	0,174239	0,128363	0	0,228965	0,239152	0,504803	0,493276	0,494505	0,498386	0,479954
21	0,446883	0,479954	0,47268	0,488246	0,468235	0,45679	0,384143	0,378663	0,384861	0,39032	0,401023	0,410891	0,468763	0,445576	0,412614	0,489787	0,405358	0,289087	0,222196	0,23035	0,228965	0	0,224905	0,498123	0,483968	0,486816	0,485403	0,486816
22	0,403746	0,44859	0,459425	0,457901	0,438984	0,418514	0,385621	0,384687	0,380414	0,390536	0,348835	0,380916	0,427013	0,366925	0,328877	0,450433	0,346014	0,319657	0,202759	0,237627	0,239152	0,224905	0	0,471415	0,452891	0,46069	0,454747	0,466826
23	0,487064	0,494939	0,49863	0,49197	0,487083	0,447165	0,470519	0,464472	0,438984	0,41891	0,441194	0,447205	0,366458	0,37932	0,375305	0,386641	0,403321	0,489882	0,475009	0,492349	0,504803	0,498123	0,471415	0	0,105163	0,247945	0,143823	0,132123
24	0,472972	0,481147	0,485976	0,478432	0,473418	0,432315	0,455519	0,446823	0,421945	0,400363	0,428291	0,431332	0,335322	0,353074	0,347335	0,361618	0,379561	0,480859	0,457841	0,481243	0,493276	0,483968	0,452891	0,105163	0	0,247894	0,144454	0,145654
25	0,478451	0,486645	0,488969	0,489483	0,48324	0,439796	0,460058	0,453489	0,427879	0,406459	0,436095	0,439714	0,354113	0,364008	0,361394	0,36441	0,385643	0,487732	0,458851	0,479492	0,494505	0,486816	0,46069	0,247945	0,247894	0	0,22159	0,20059
26	0,497427	0,503218	0,507314	0,511	0,5003	0,461992	0,466826	0,450293	0,447808	0,426395	0,453349	0,468978	0,375239	0,402216	0,389543	0,399872	0,421032	0,486091	0,459168	0,486435	0,498386	0,485403	0,464747	0,143823	0,144454	0,22159	0	0,208765
27	0,486645	0,513218	0,509428	0,515495	0,505418	0,464472	0,439796	0,464472	0,44626	0,412595	0,453885	0,461992	0,366458	0,418285	0,375305	0,386641	0,403321	0,475009	0,445576	0,489787	0,479954	0,486816	0,466826	0,132123	0,145654	0,20059	0,208765	0,203654

Rys. 7. Wyniki obliczonego współczynnika różnicy Rogera– Tanimoto dla zbioru małych krzywych.

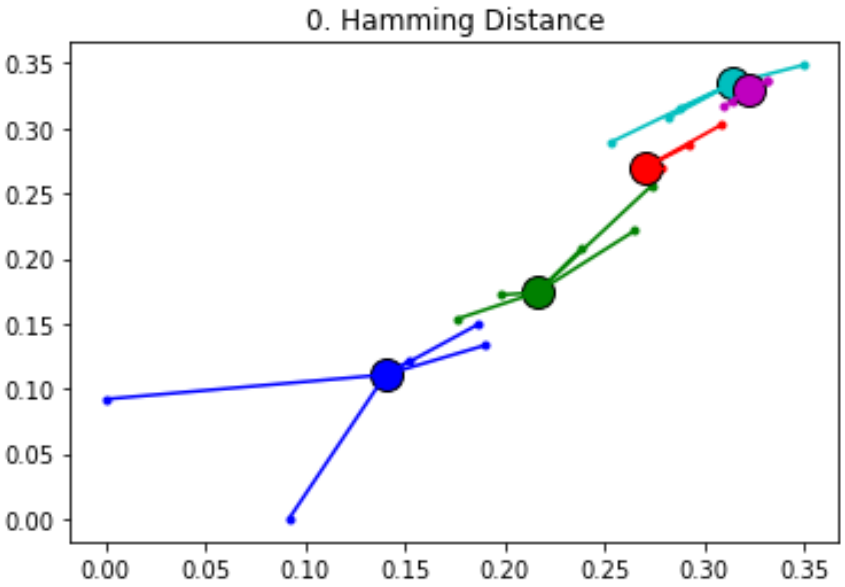
Po obliczeniu macierzy podobieństwa, skuteczność każdej z miar odległości/podobieństwa oceniania była za pomocą algorytmów klastrujących. Z racji faktu, że przeprowadzany proces analizy jest procesem nienadzrowanym, do klastrowania wykorzystane zostały algorytmy, które same podejmują decyzję o ilości skupień. Reprezentantami powyższej grupy algorytmów są:

- Propagacja powinowactwa.
- Metoda pojedynczego wiązania.
- Metoda pełnego wiązania.
- Metoda środkowego wiązania.

W dalszej części tego rozdziału znajdują się wyniki przeprowadzonej analizy.

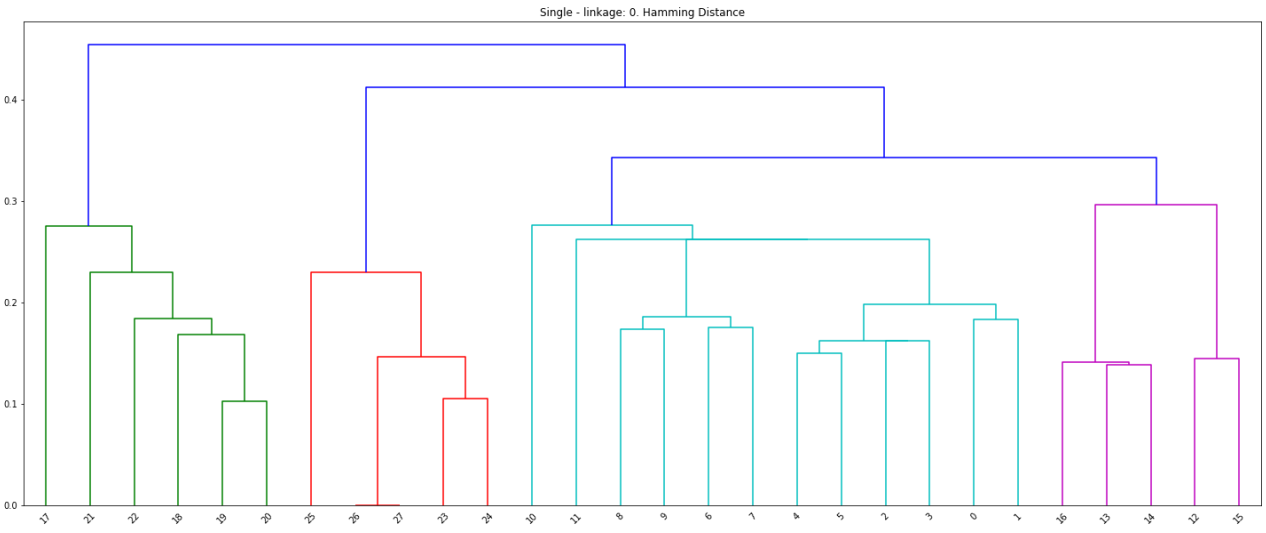
Analiza zbioru krzywych mniejszych:

1. Odległość Hamminga:
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4]
```

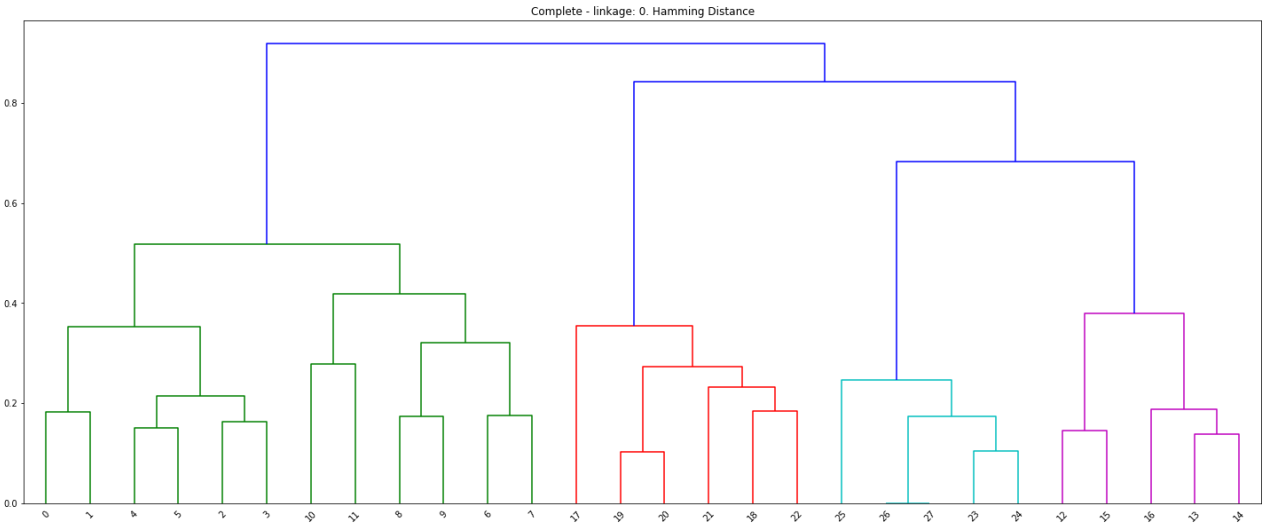
        ii. Poprawność: Analiza skupień przeprowadzona poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania

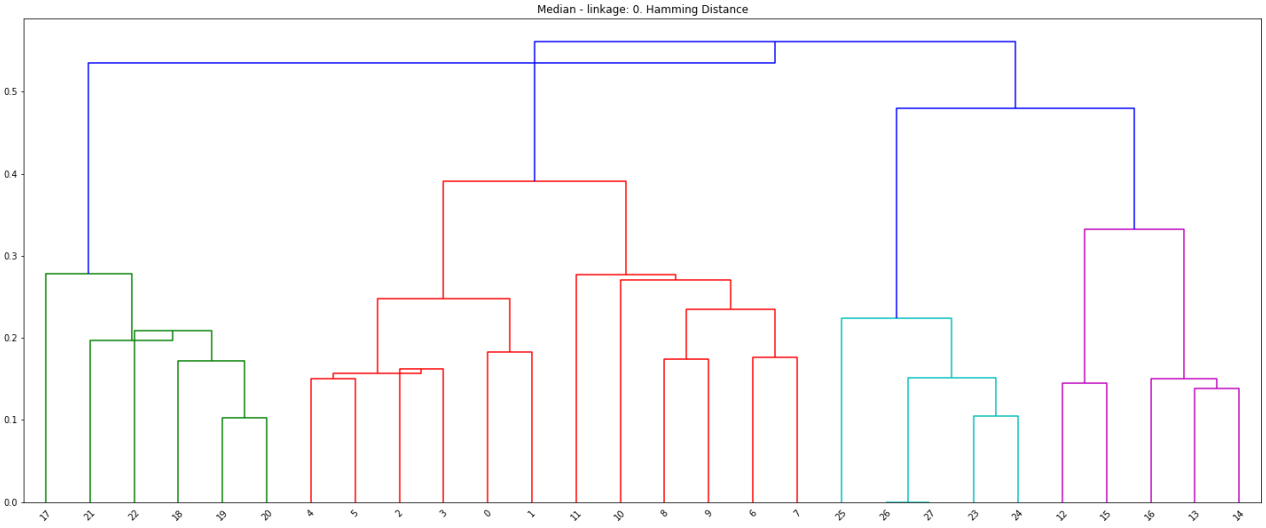
i. Dendrogram:



ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

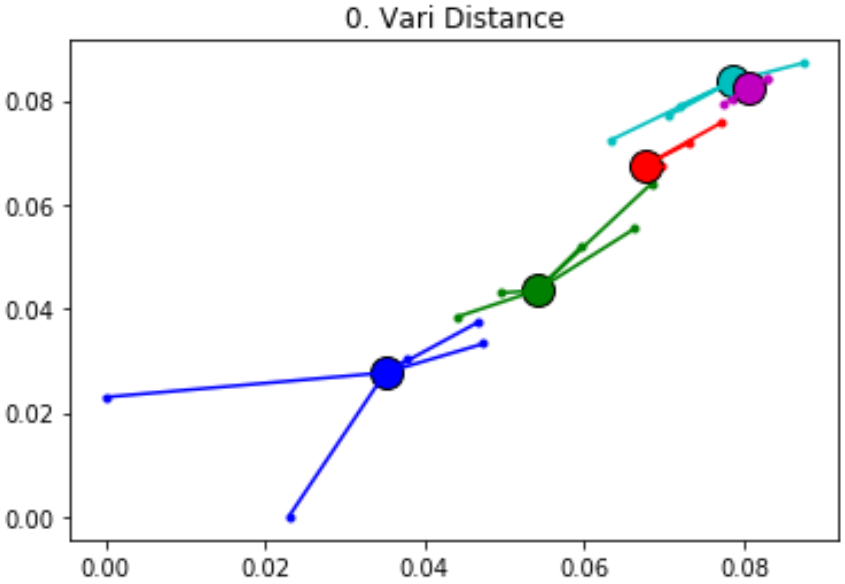
d. Metoda środkowego wiązania:

i. Dendrogram:



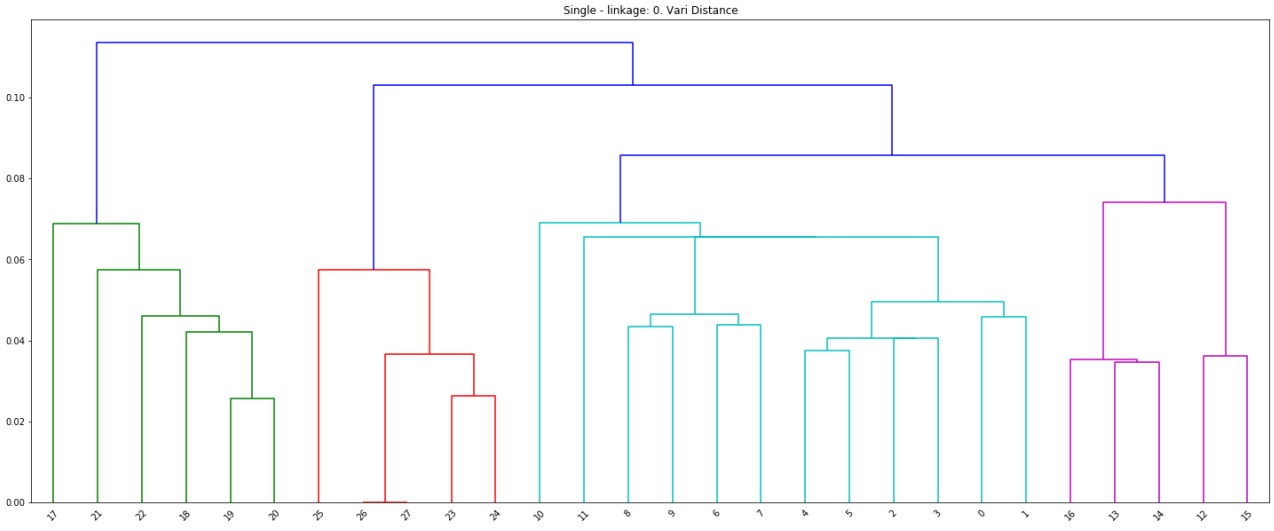
ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

2. Odległość Vari’ego:
    a. Propagacja powinowactwa:
        i. Wizualizacja:



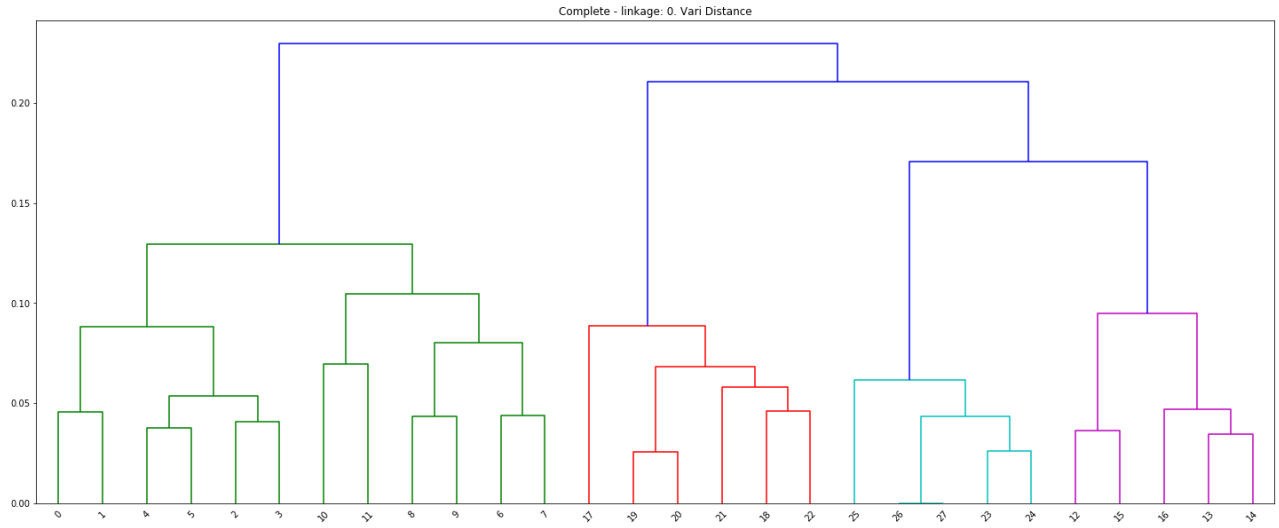

```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4]
```

        ii. Poprawność: Analiza skupień przeprowadzona poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



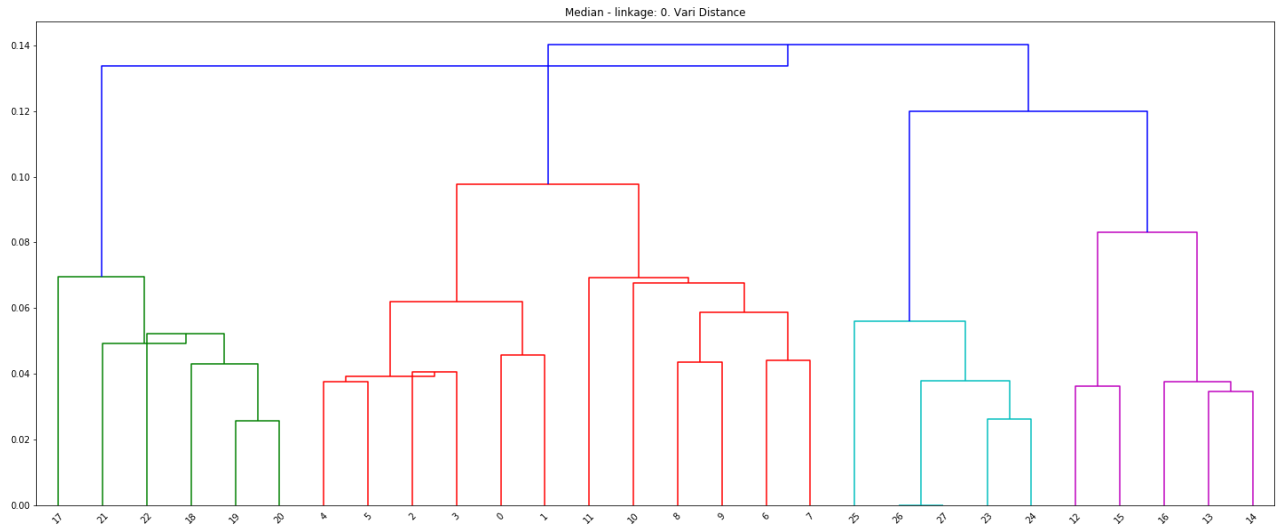
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



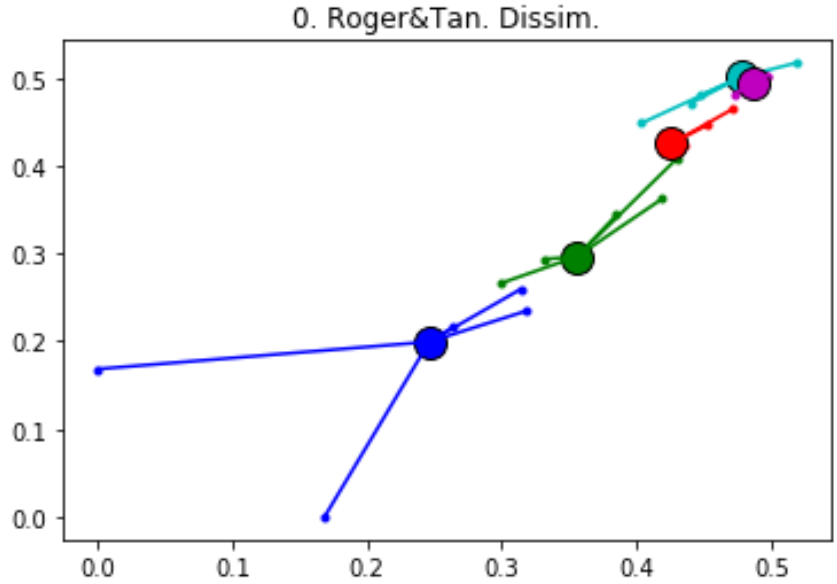
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



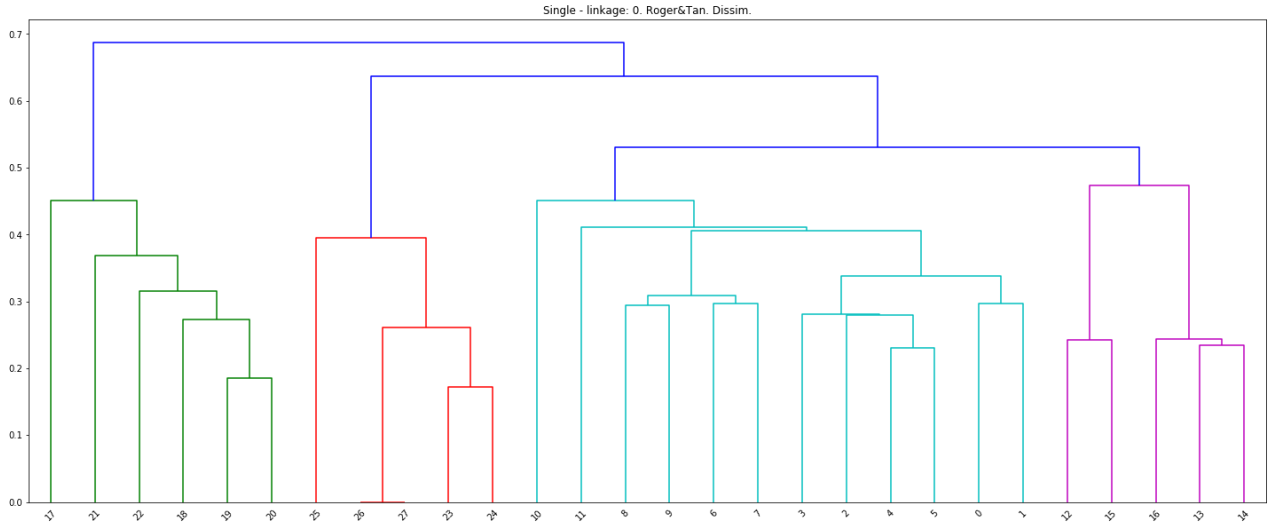
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

3. Współczynnik różnicy Rogera – Tanimoto:
    a. Propagacja powinowactwa:
        i. Wizualizacja:



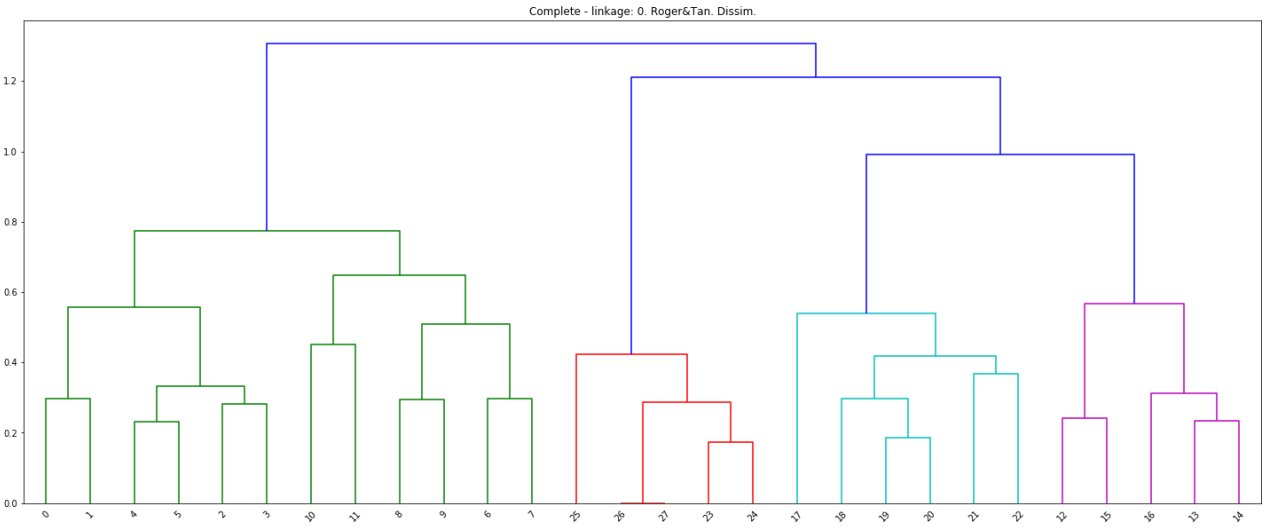

```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4]
```

        ii. Poprawność: Analiza skupień przeprowadzona poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



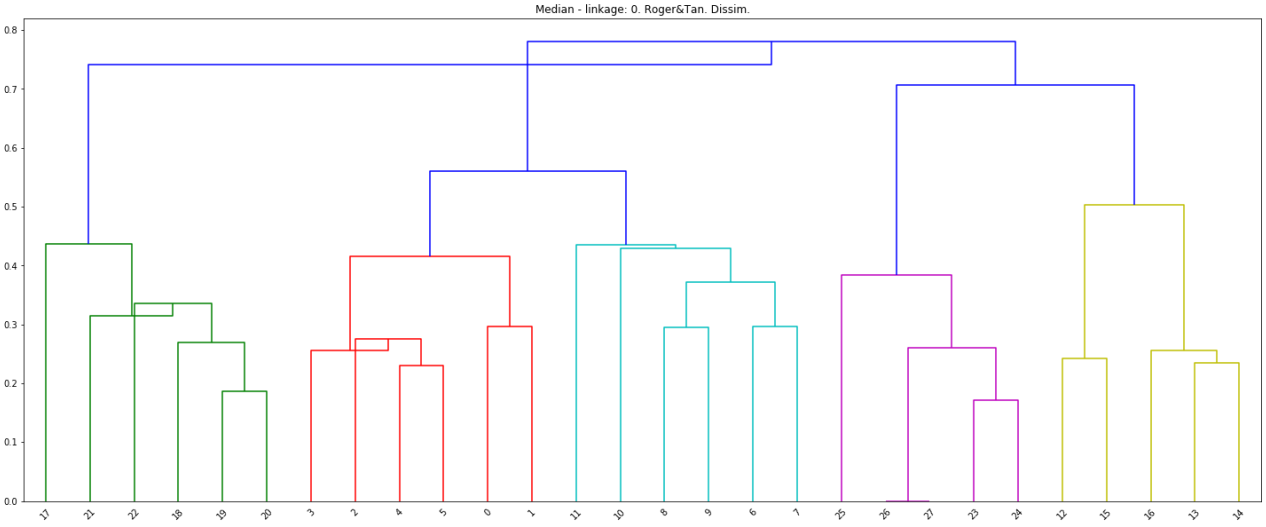
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
   i. Dendrogram:



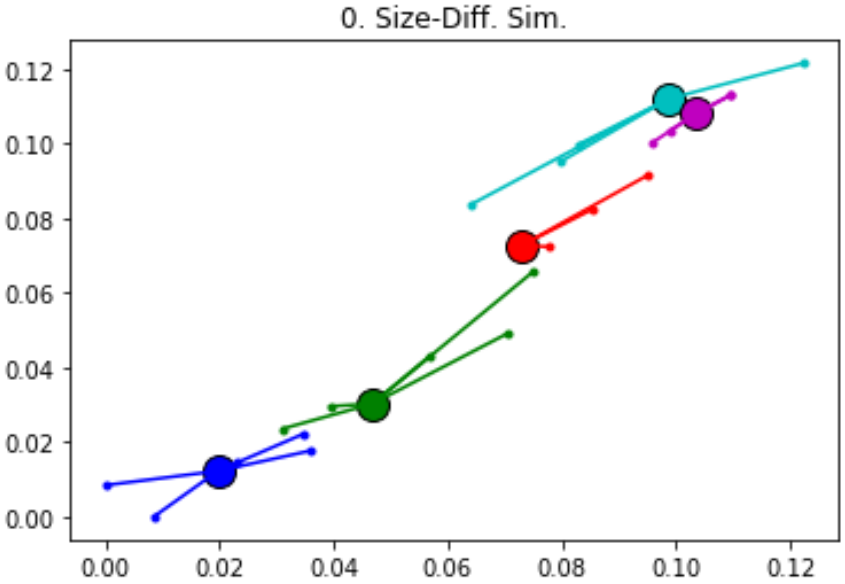
   ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
   i. Dendrogram:



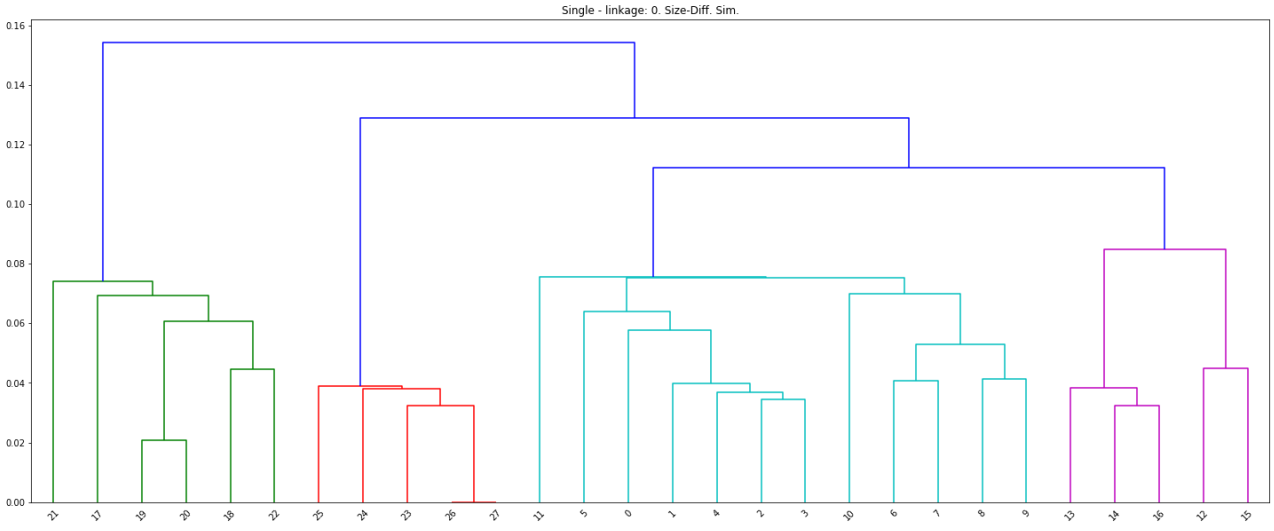
   ii. Poprawność: Analiza skupień przeprowadzona poprawnie.

4. Odległość różnicy rozmiaru:
    a. Propagacja powinowactwa:
        i. Wizualizacja:



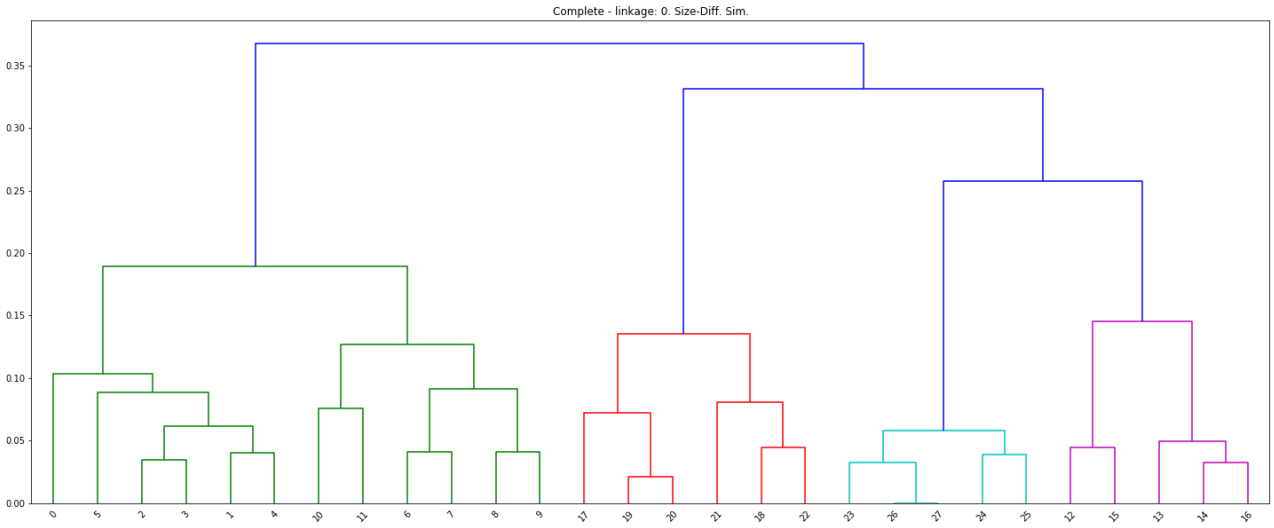

```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4]
```

        ii. Poprawność: Analiza skupień przeprowadzona poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



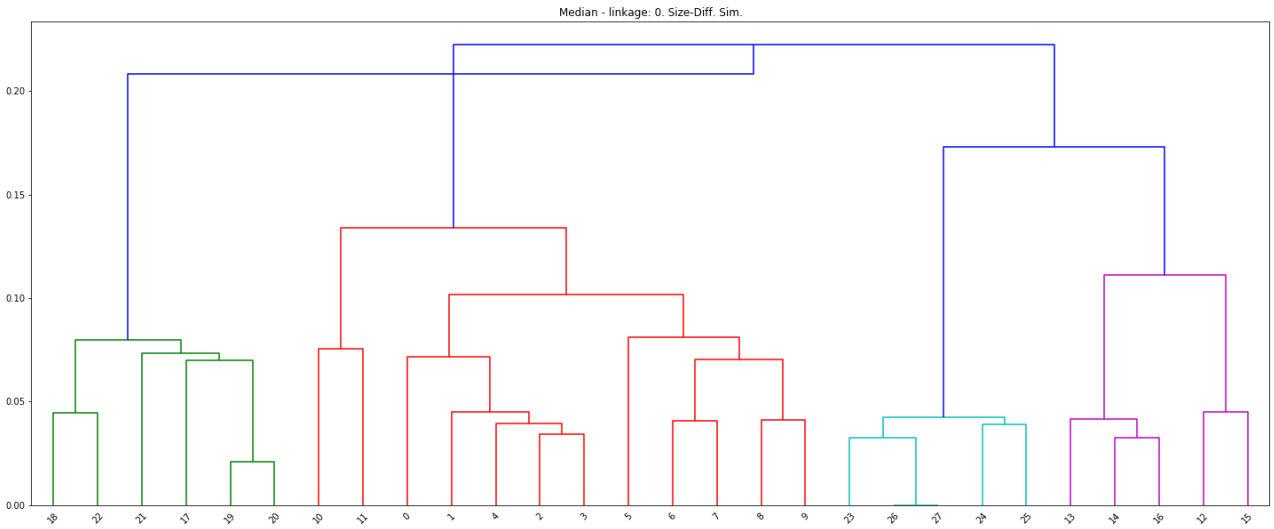
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



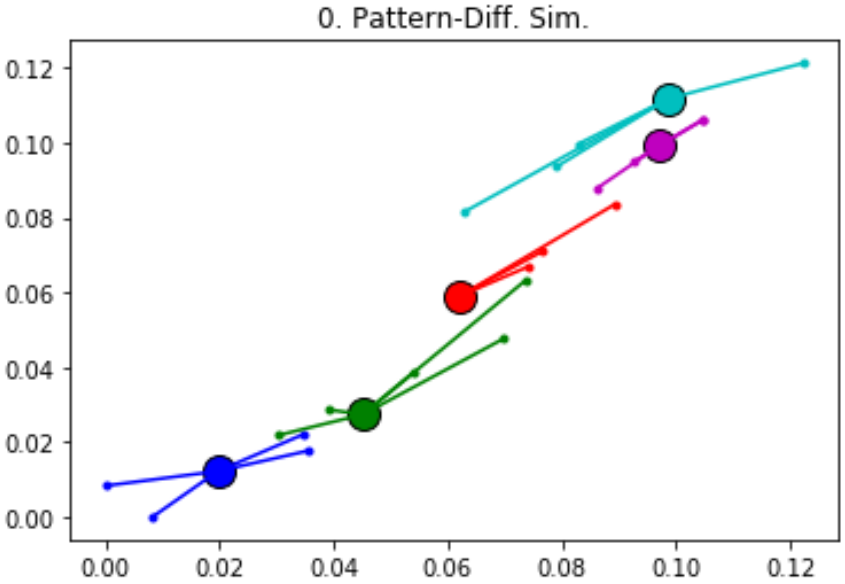
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



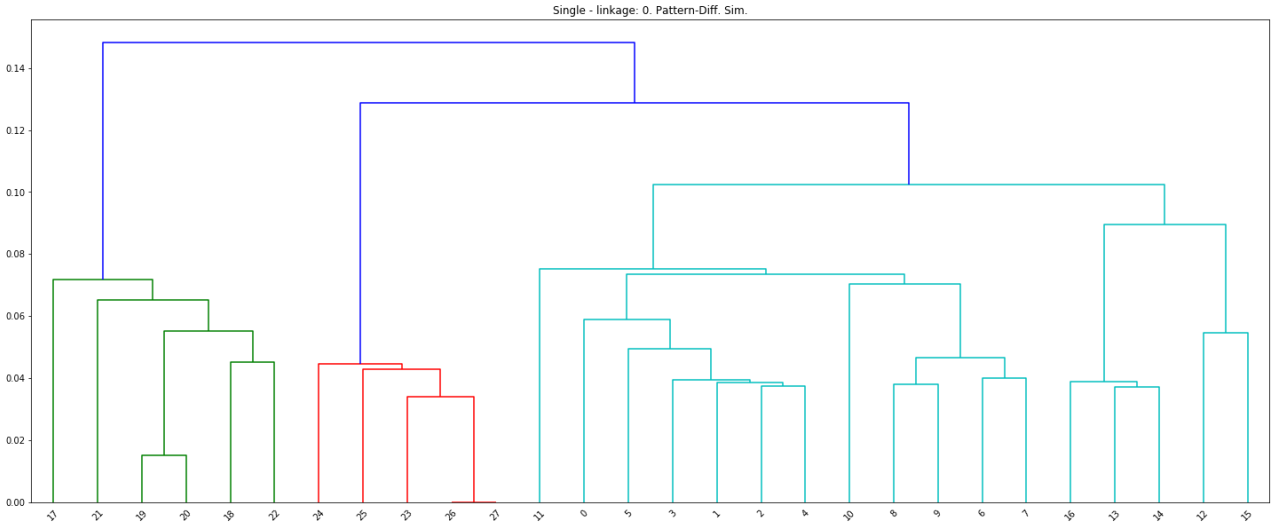
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

5. Odległość różnicy wzorca:
    a. Propagacja powinowactwa:
        i. Wizualizacja:



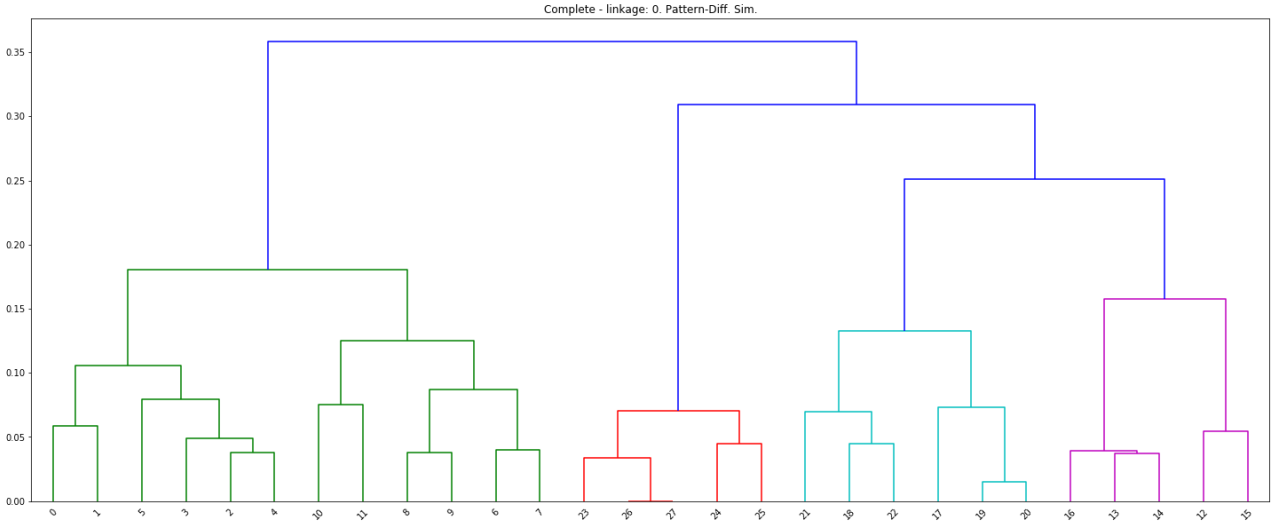

```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4]
```

        ii. Poprawność: Analiza skupień przeprowadzona poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



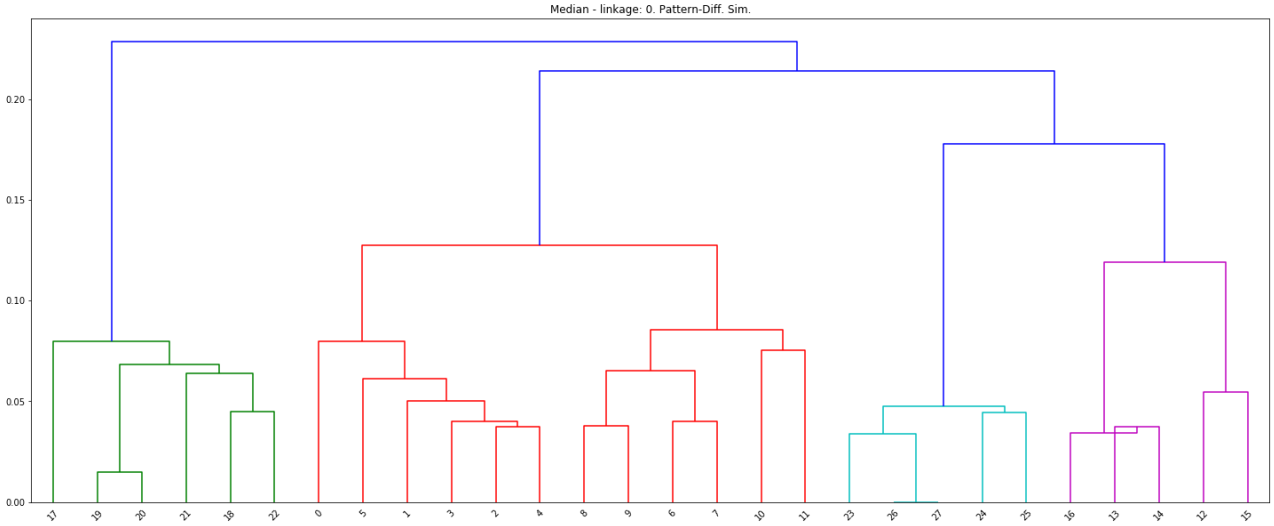
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Pozostałe dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



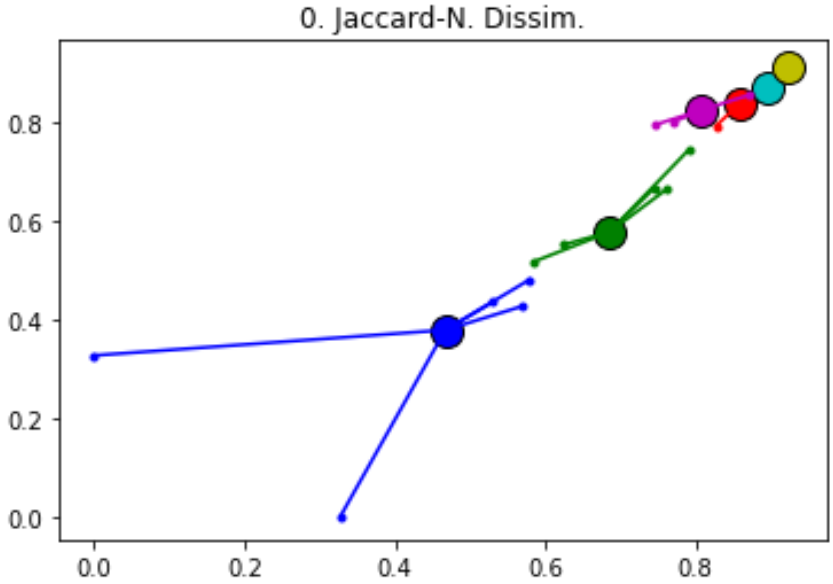
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



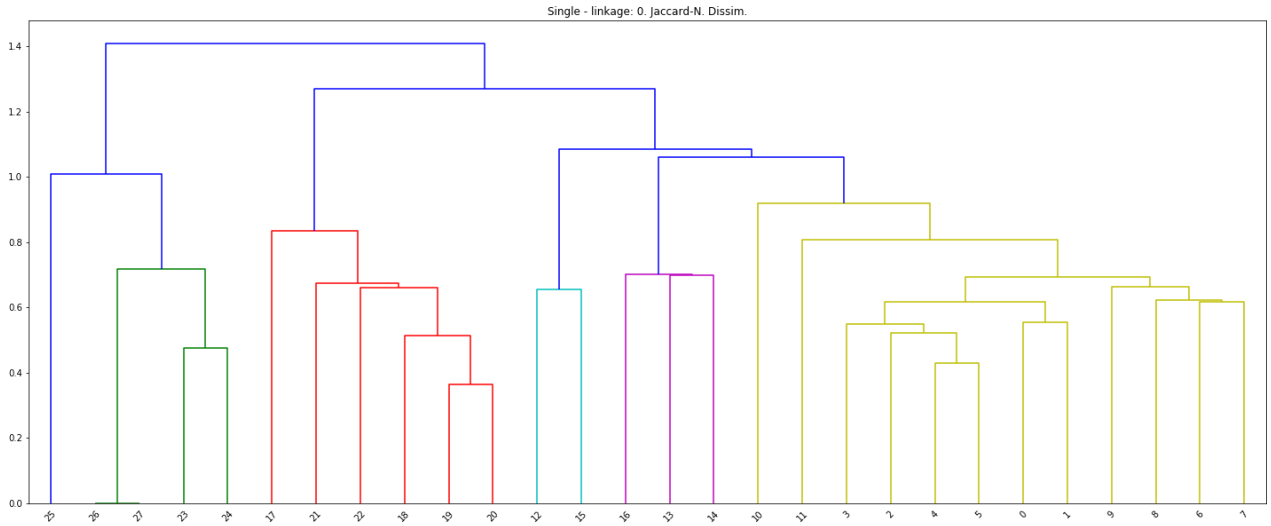
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

6. Współczynnik różnicy Jaccarta – Needhama
    a. Propagacja powinowactwa:
        i. Wizualizacja:



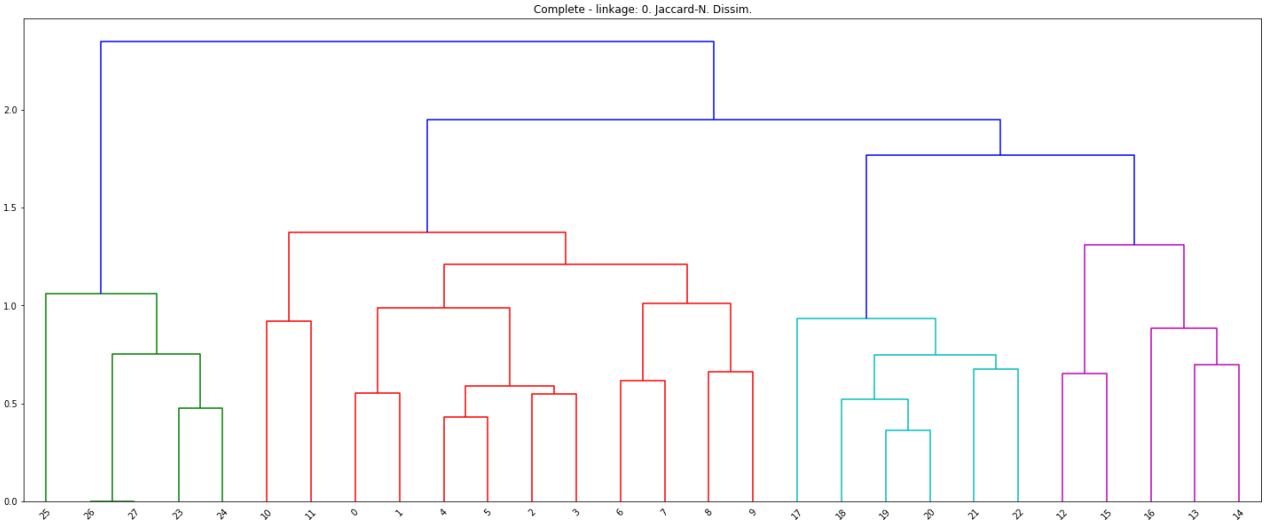

```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 3 2 2 3 2 4 4 4 4 4 4 5 5 5 5 5]
```

        ii. Poprawność: Źle rozpoznana ilość klastrów. Trzeci klaster rozłączony na dwa mniejsze. Pozostałe grupy rozpoznane poprawnie.
    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



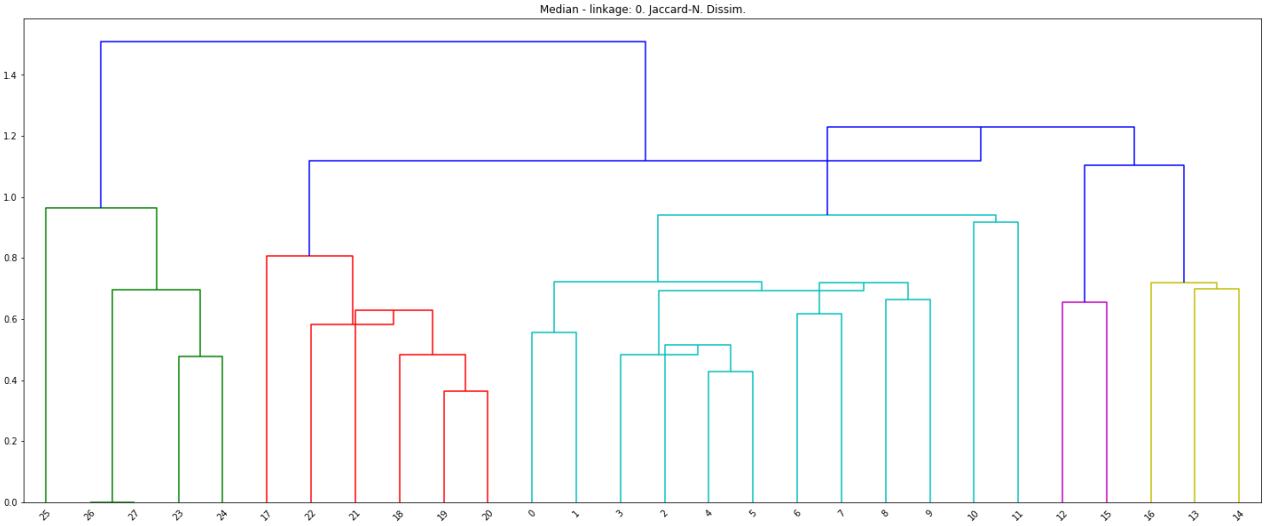
        ii. Poprawność: Źle rozpoznana ilość klastrów, wiele elementów źle dopasowanych.

c. Metoda pełnego wiązania:
    i. Dendrogram:



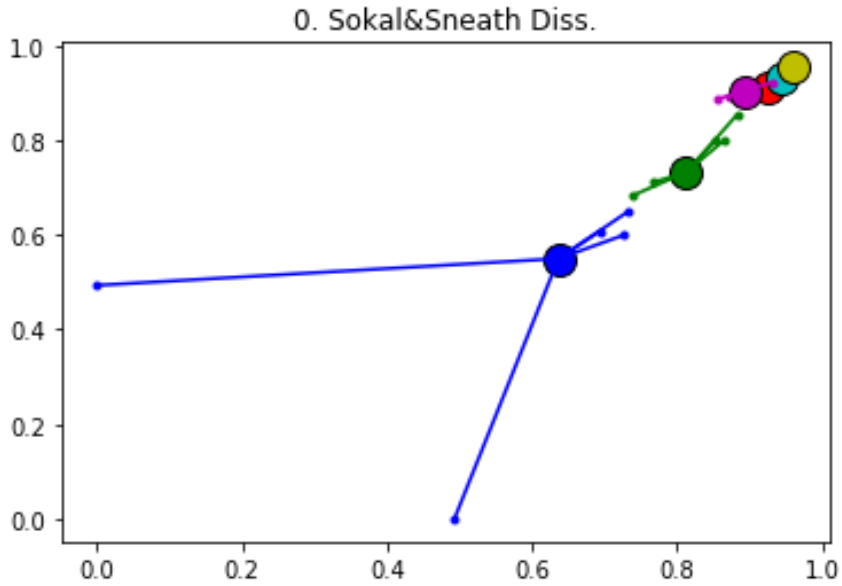
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Klaster trzeci rozdzielony na dwie osobne grupy. Pozostałe dwie grupy rozpoznane poprawnie.

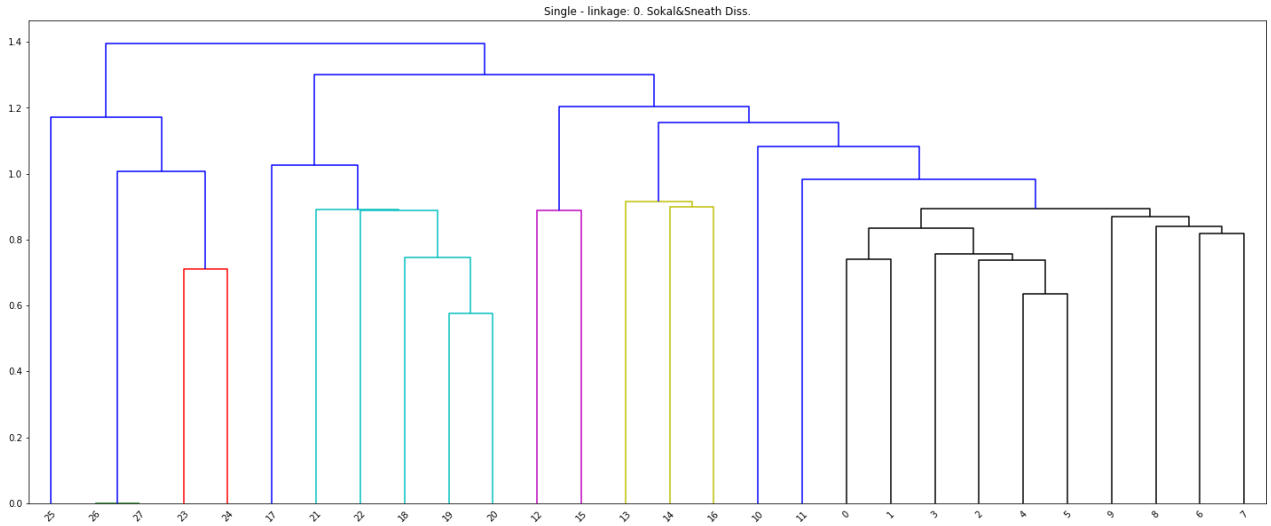
7. Współczynnik różnicy Sokala – Sneatha:
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 3 2 2 3 2 4 4 4 4 4 4 5 5 5 5 5]
```

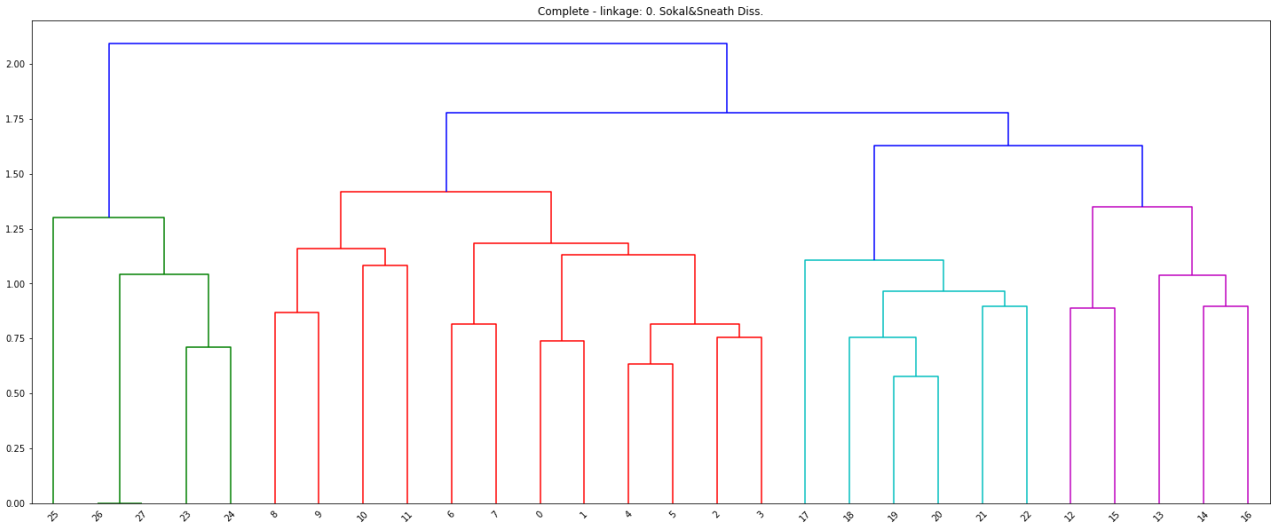
        ii. Poprawność: Źle rozpoznana ilość klastrów. Trzeci klaster rozłączony na dwa mniejsze. Pozostałe grupy rozpoznane poprawnie.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



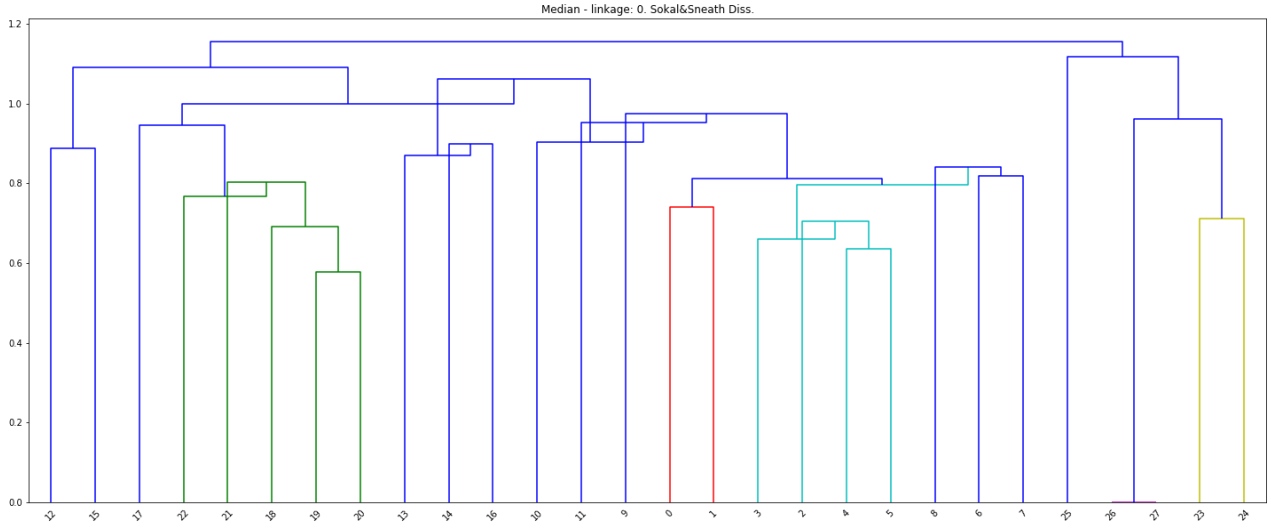
        ii. Poprawność: Źle rozpoznana ilość klastrów. Wiele elementów nie zostało przyporządkowanych do żadnej z grup.

c. Metoda pełnego wiązania:
    i. Dendrogram:



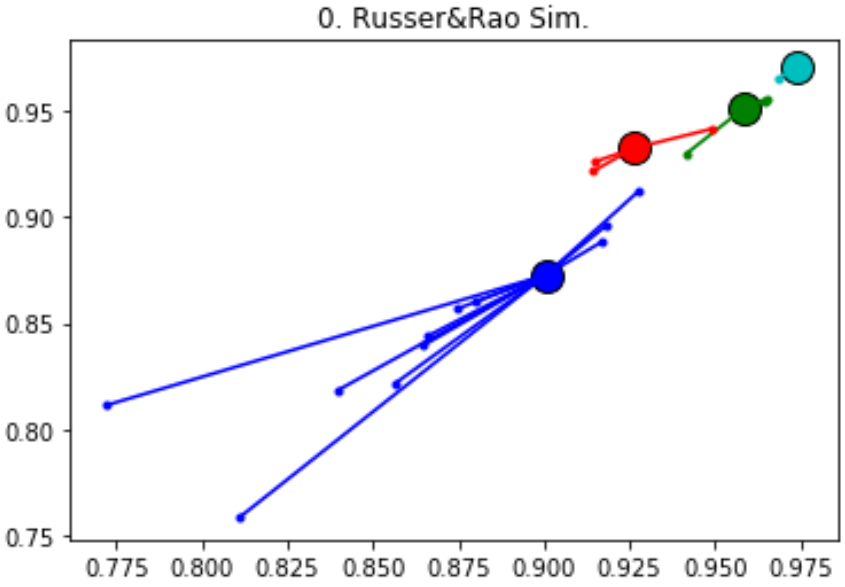
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Źle rozpoznana ilość klastrów. Wiele elementów nie zostało przyporządkowanych do żadnej z grup.

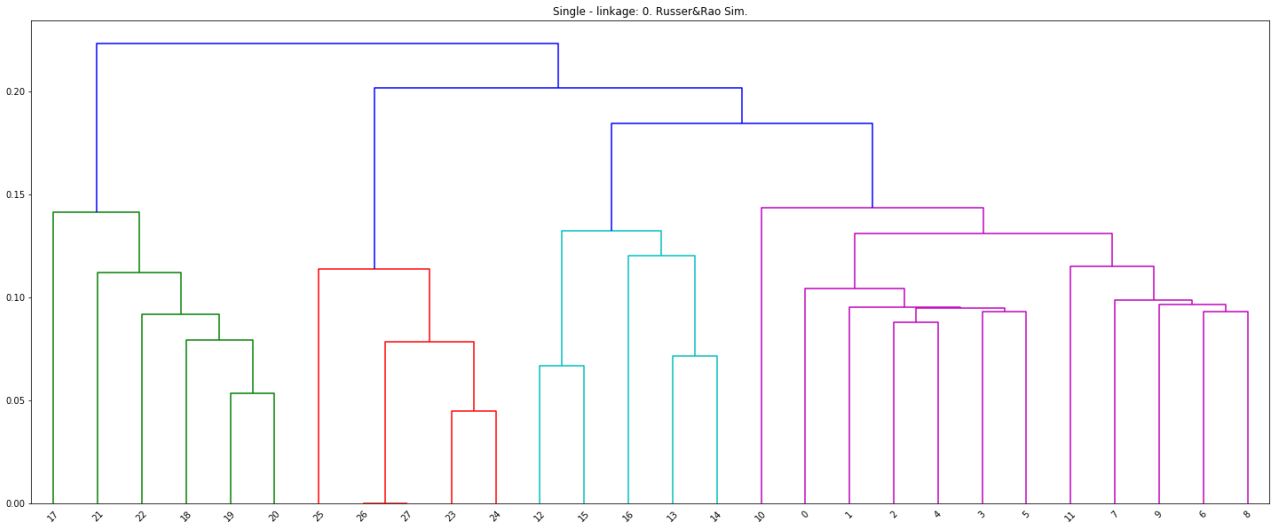
8. Współczynnik różnicy Russera – Rao:
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3]
```

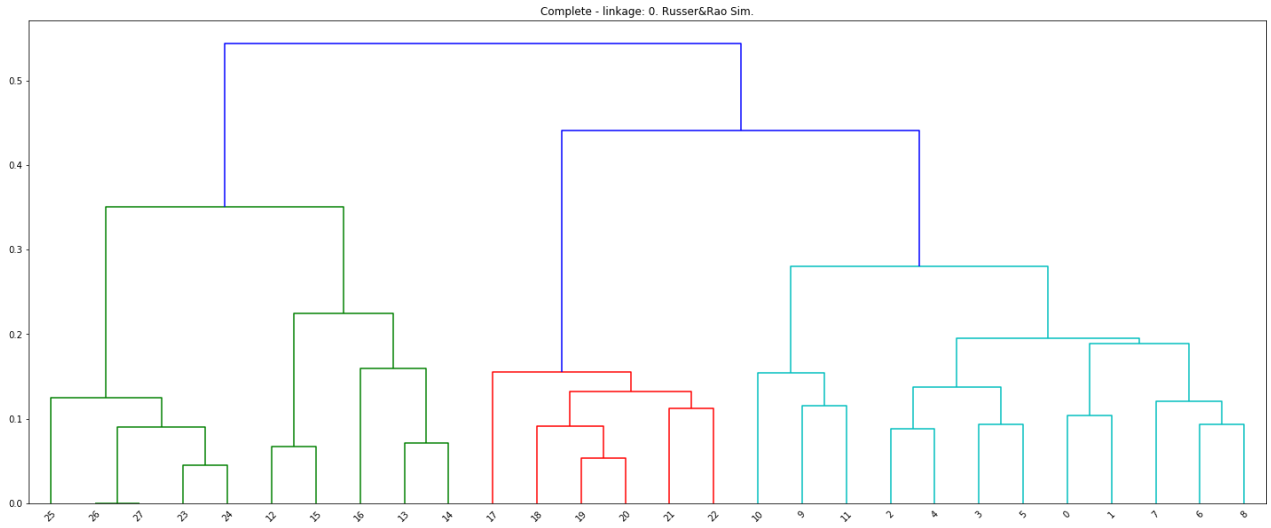
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



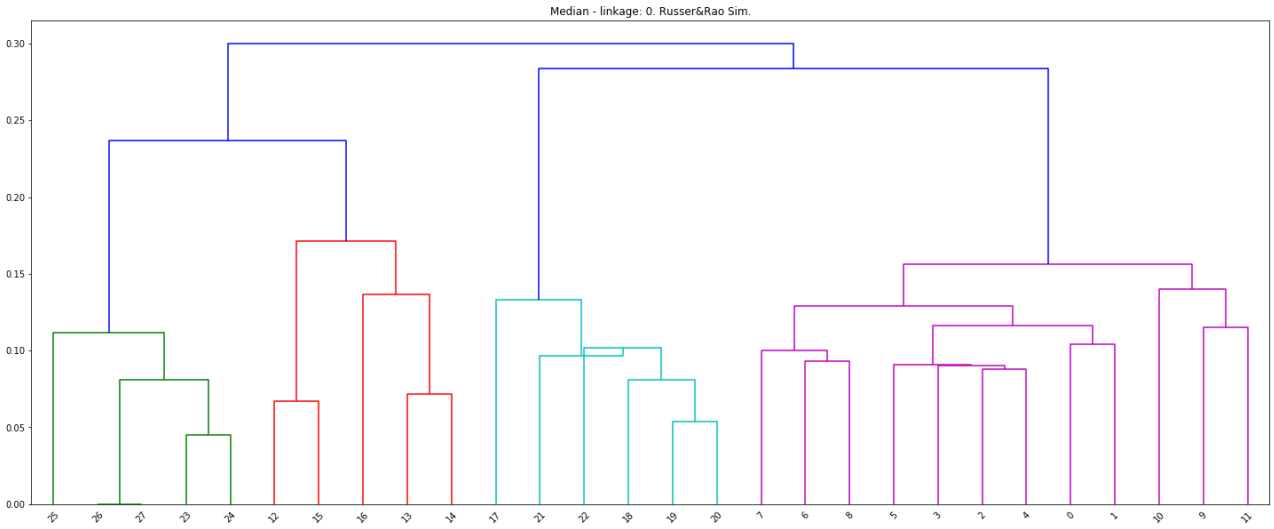
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Pozostałe dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



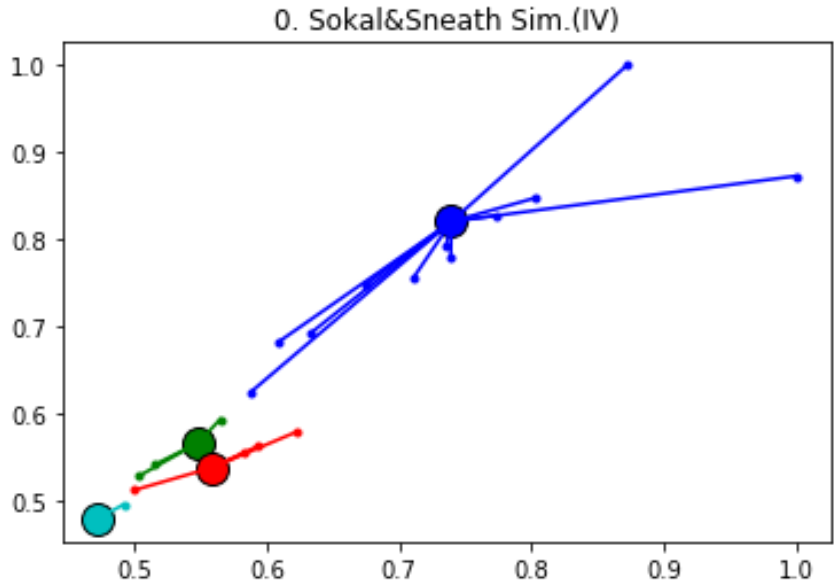
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim, klaster trzeci z klastrem piątym. Tylko grupa czwarta rozpoznana poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

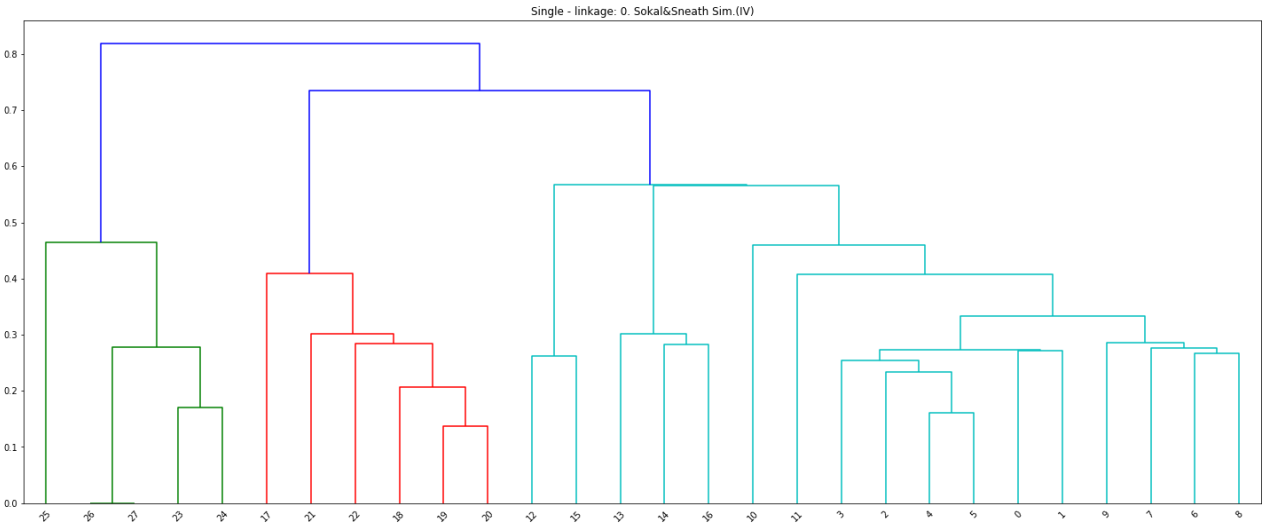
9. Współczynnik podobieństwa Sokala – Sneatha (IV):
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3]
```

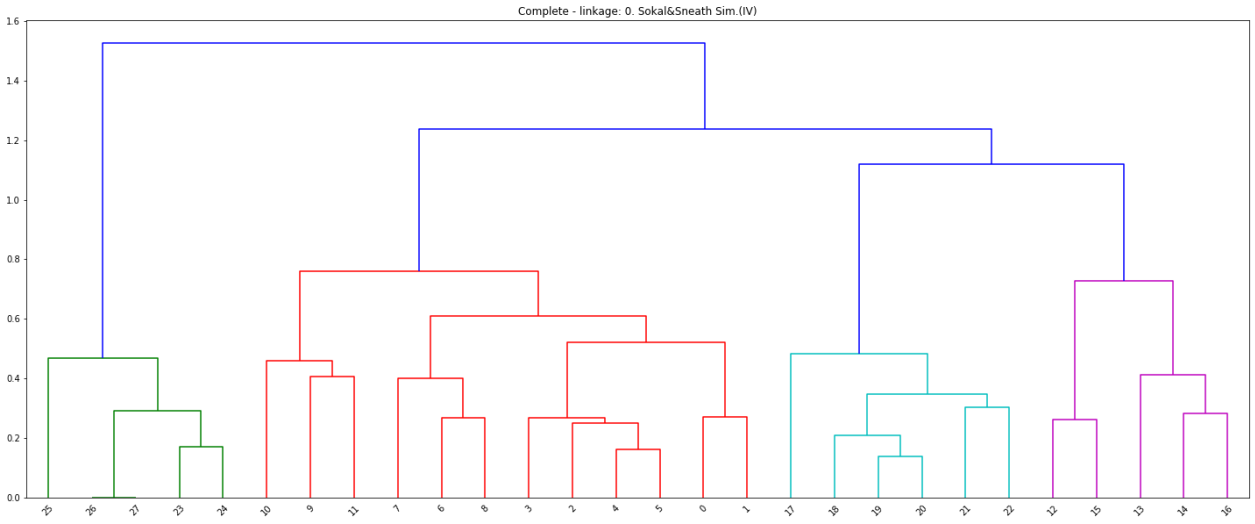
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



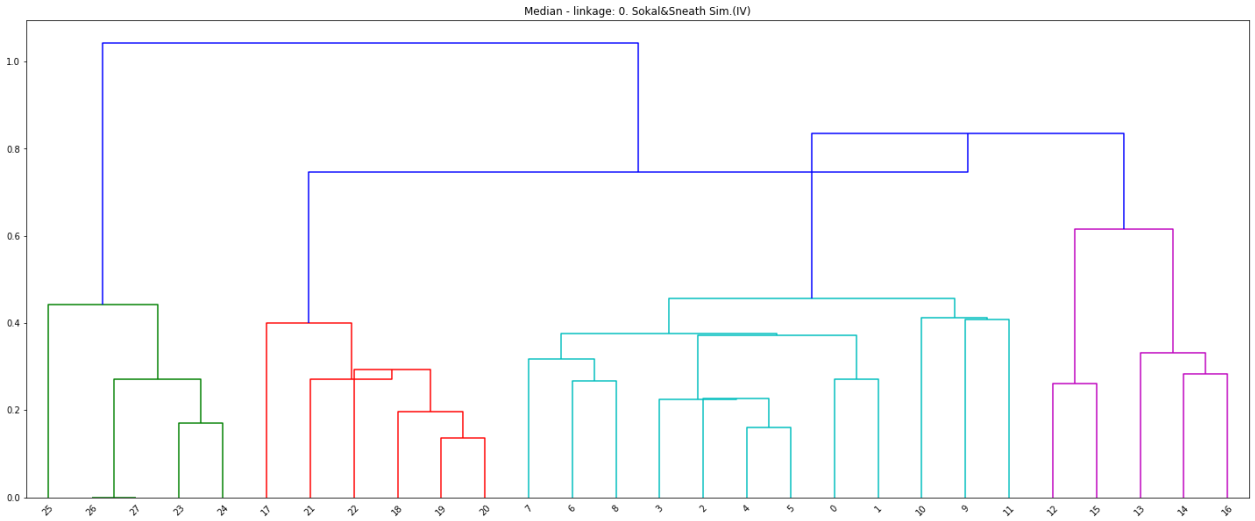
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Pozostałe dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



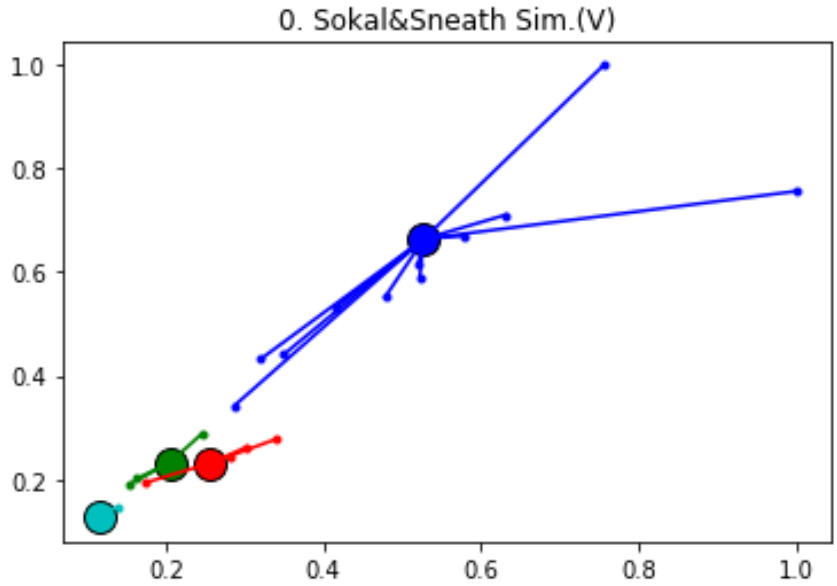
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

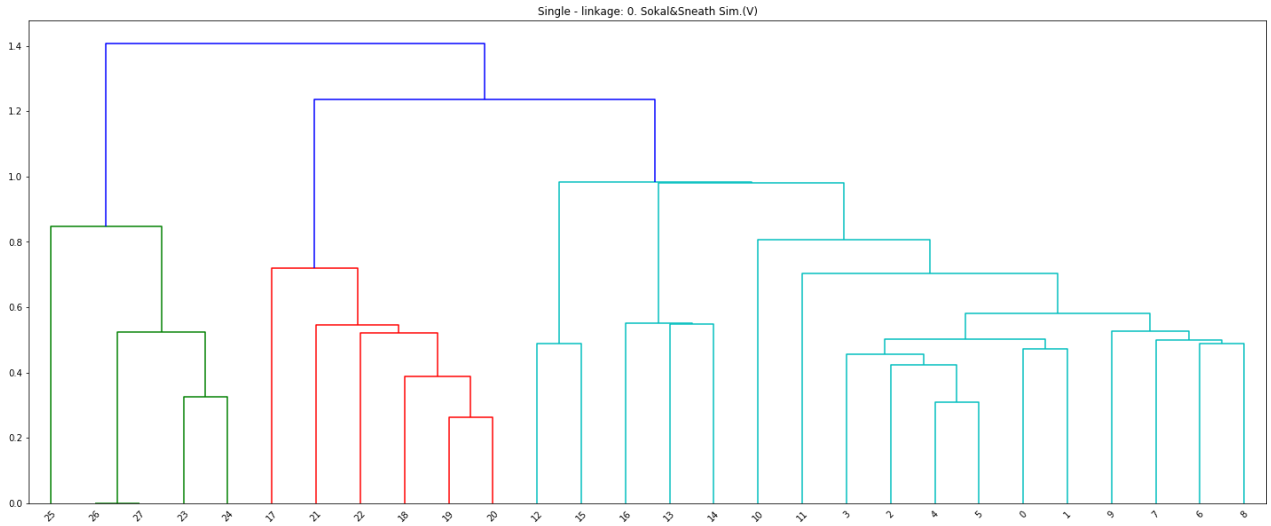
10. Współczynnik podobieństwa Sokala – Sneatha (V):
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3]
```

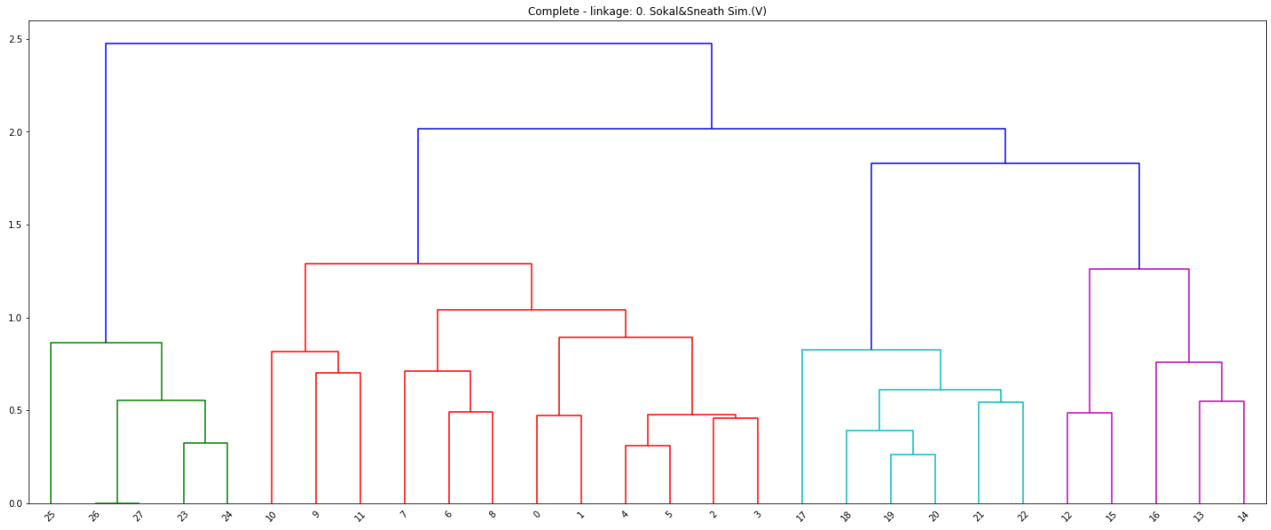
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



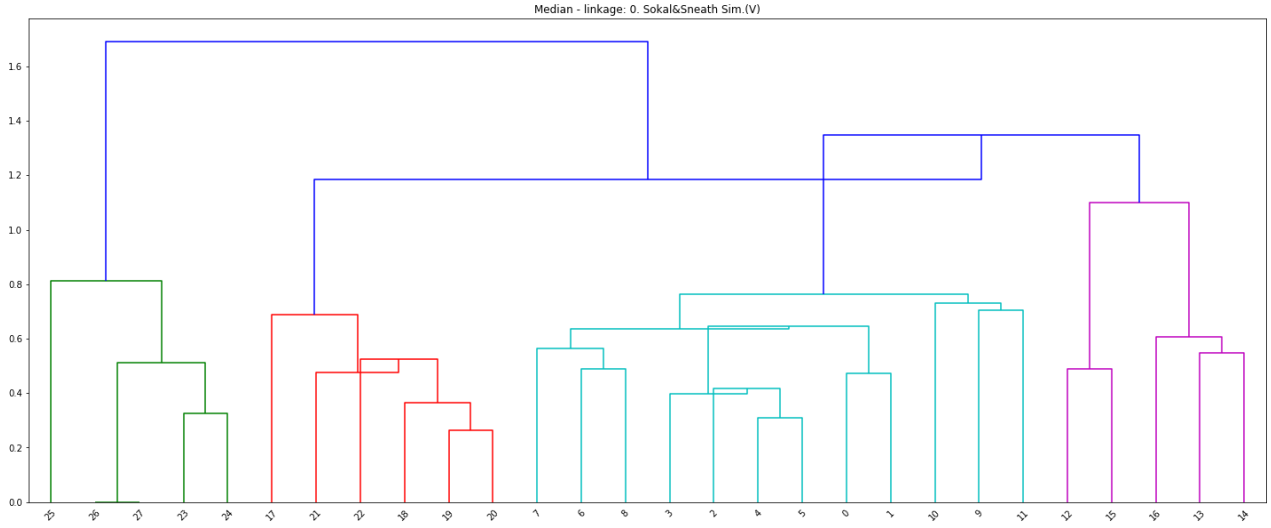
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Pozostałe dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:

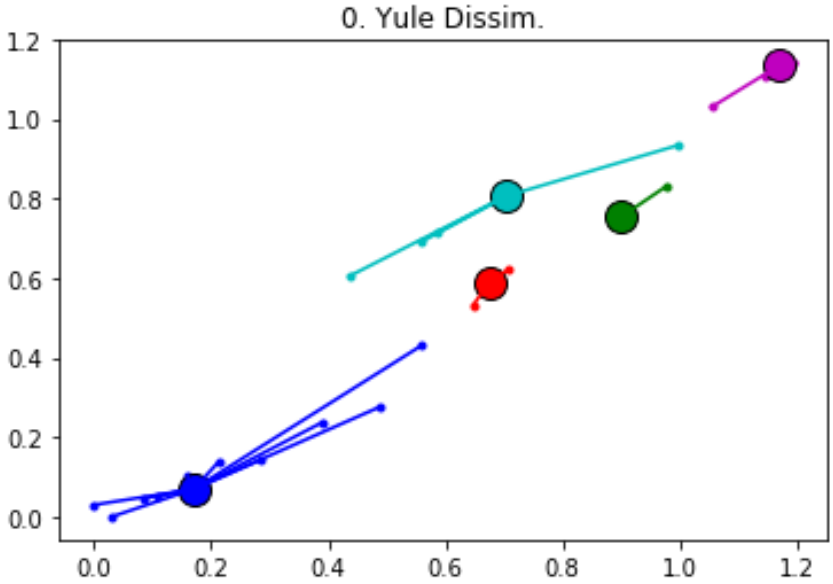


    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

11. Współczynnik różnicy Yule'a:

a. Propagacja powinowactwa:

i. Wizualizacja:

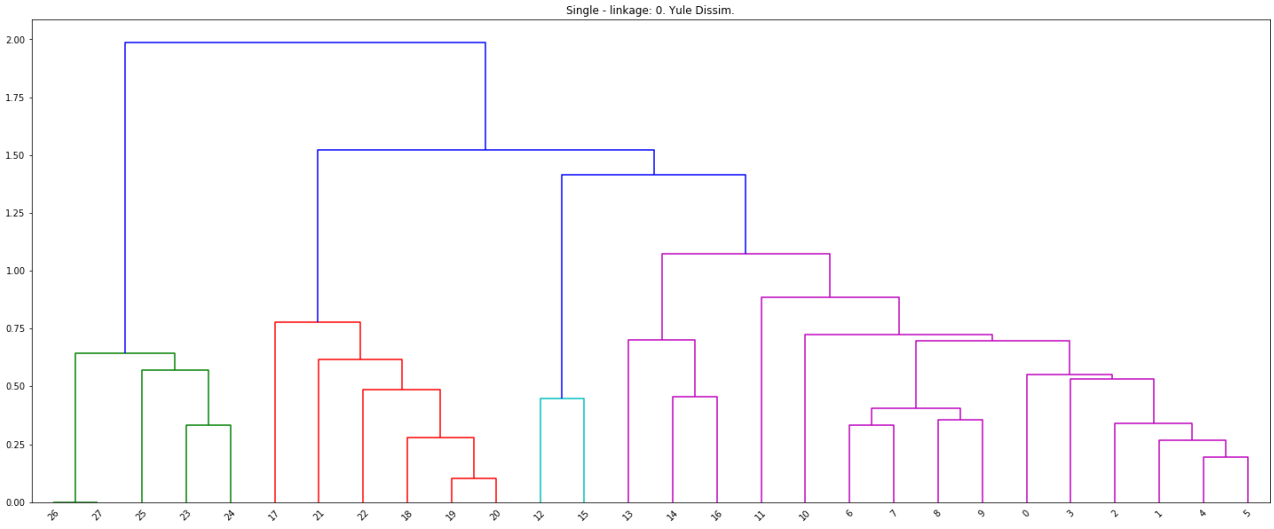



```
Result: [0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 1 2 3 3 3 3 3 3 4 4 4 4 4]
```

ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Trzecia grupa rozłączona na dwie mniejsze. Ostatnie dwie grupy rozpoznane poprawnie.

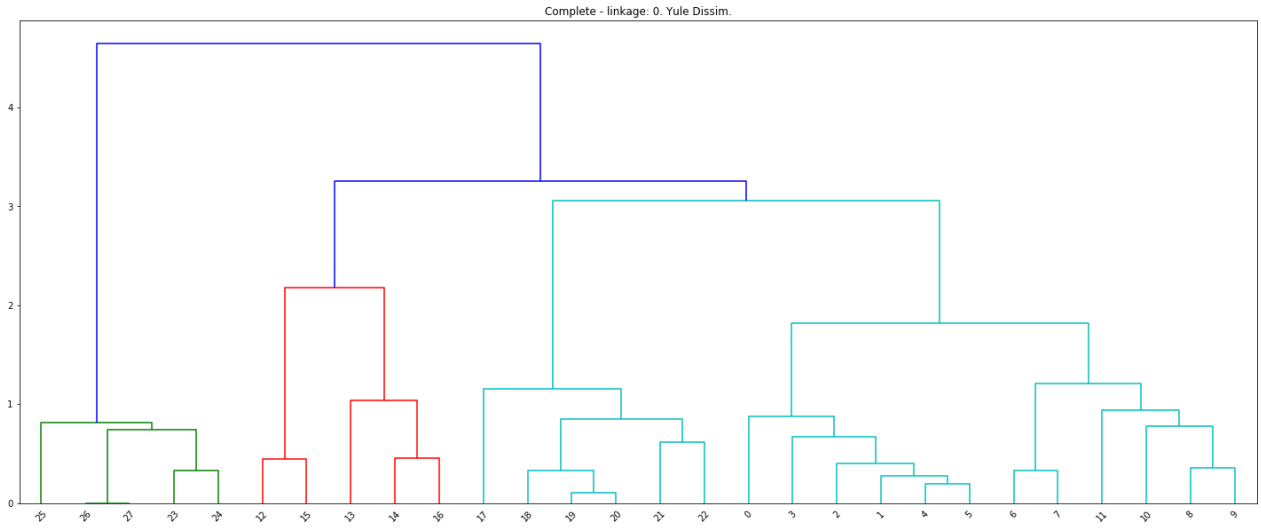
b. Metoda pojedynczego wiązania:

i. Dendrogram:



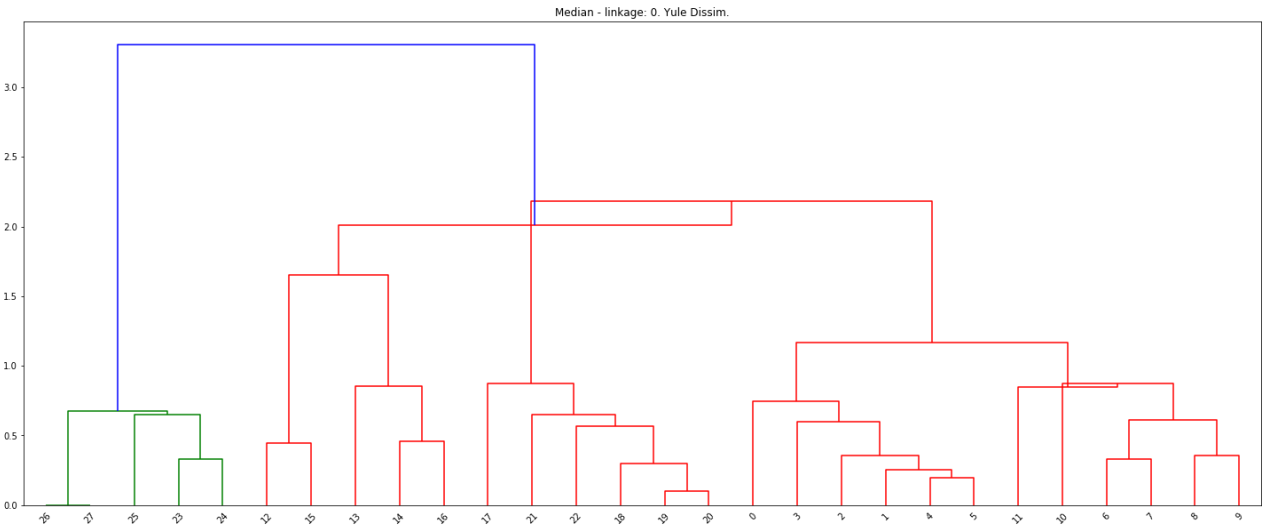
ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Wydzielona błędna dodatkowa grupa. Ostatnie dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i trzecim. Pozostałe dwie grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:

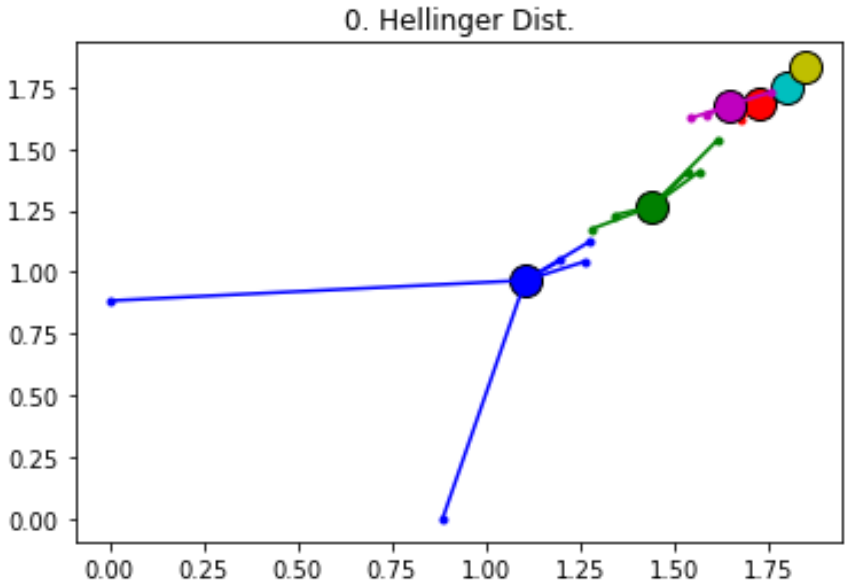


    ii. Poprawność: Źle rozpoznana ilość klastrów. Tylko ostatnia grupa wydzielona poprawnie.

12. Odległość Hellingera:

a. Propagacja powinowactwa:

i. Wizualizacja:

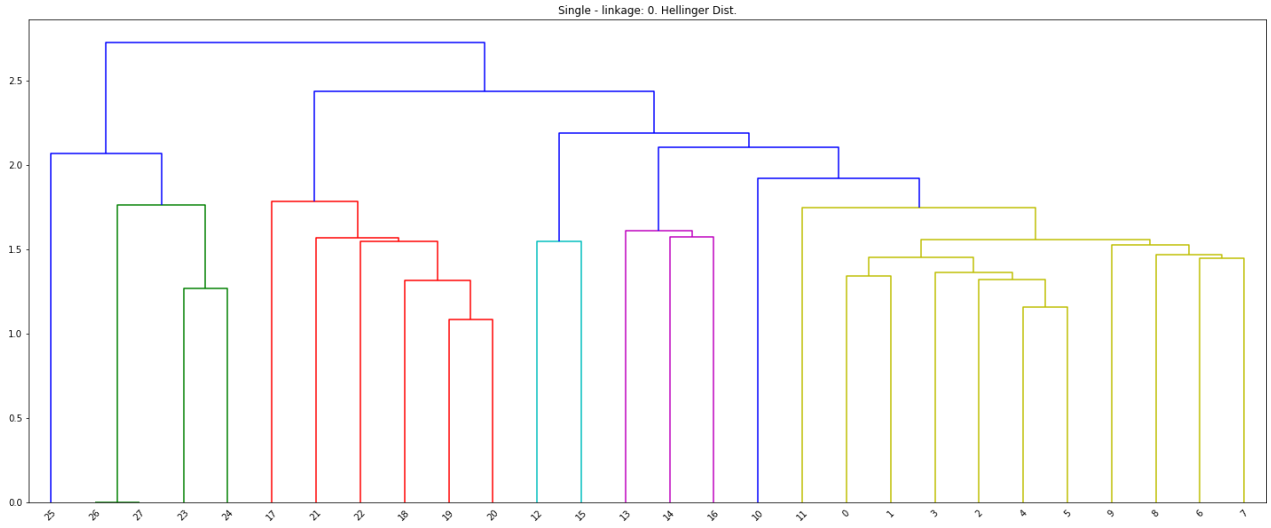



```
Result: [0 0 0 0 0 0 1 1 1 1 1 1 3 2 2 3 2 4 4 4 4 4 4 5 5 5 5 5]
```

ii. Poprawność: Źle rozpoznana ilość klastrów. Trzeci klaster rozłączony na dwa mniejsze. Pozostałe grupy rozpoznane poprawnie.

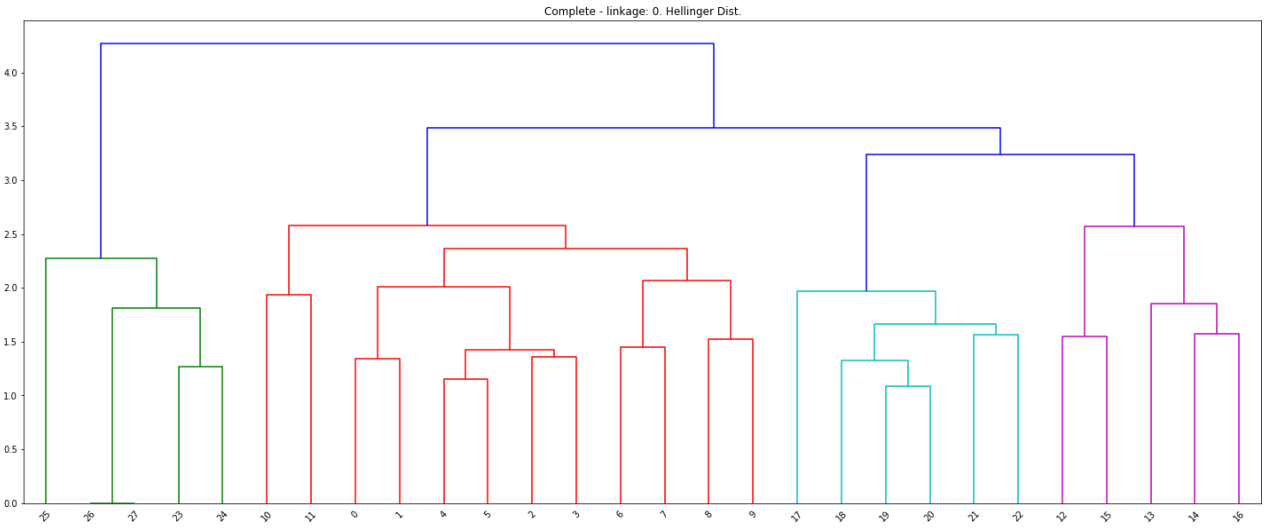
b. Metoda pojedynczego wiązania:

i. Dendrogram:



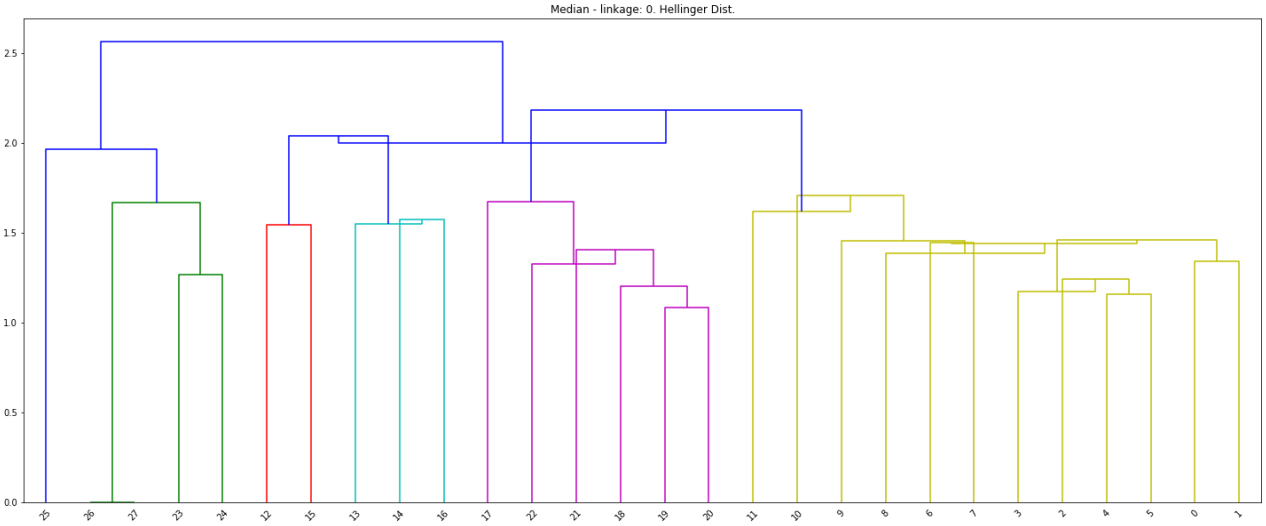
ii. Poprawność: Źle rozpoznana ilość klastrów. Wiele elementów nie zostało przyporządkowanych do żadnej z grup.

c. Metoda pełnego wiązania:
    i. Dendrogram:



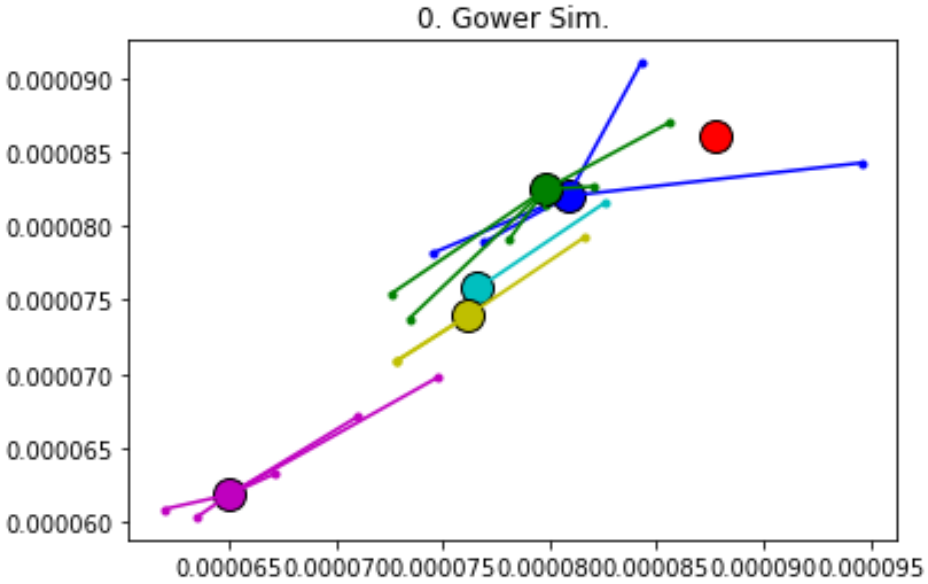
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Pozostałe grupy rozpoznane poprawnie.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Błędne przyporządkowanie elementów do grup.

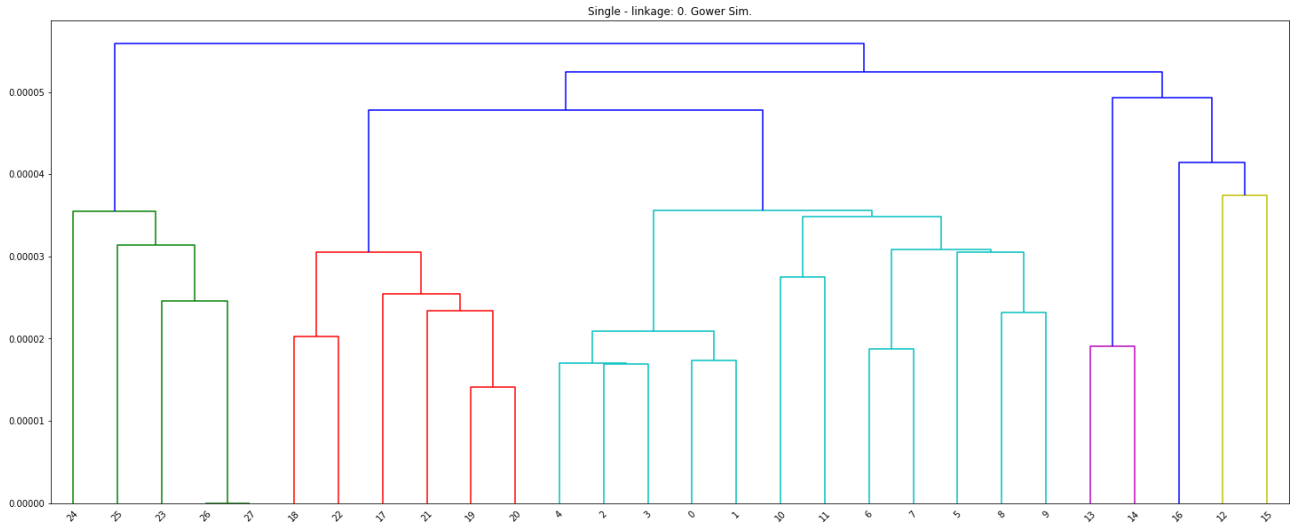
13. Współczynnik podobieństwa Gowera:
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 1 1 1 1 1 1 1 3 2 2 3 3 4 4 4 4 4 4 5 5 5 5 5]
```

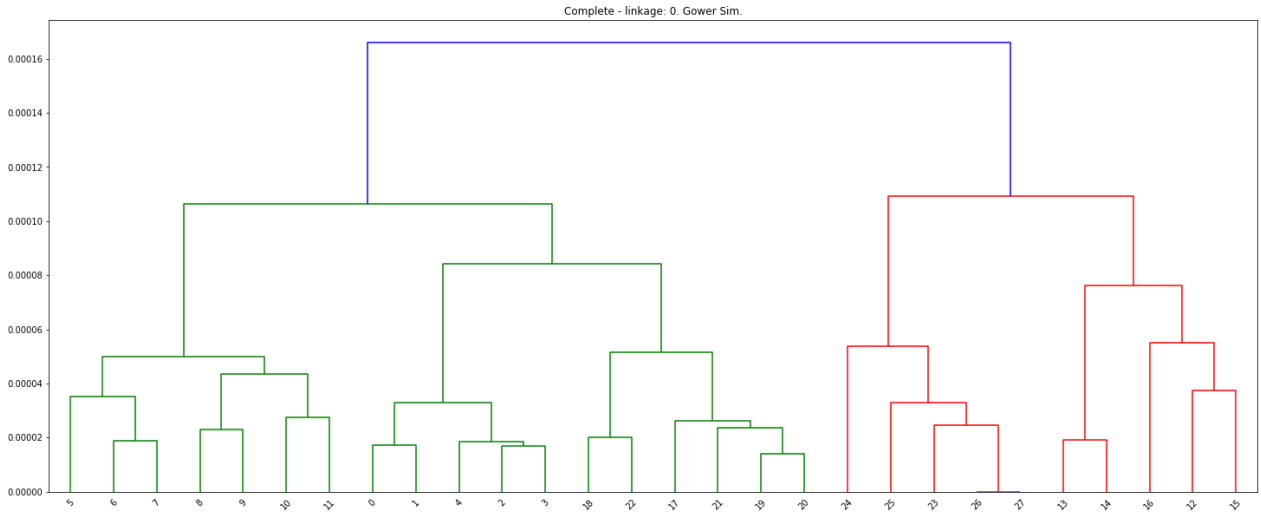
        ii. Poprawność: Źle rozpoznana ilość klastrów. Trzeci klaster rozłączony na dwa mniejsze. Pozostałe grupy rozpoznane poprawnie.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



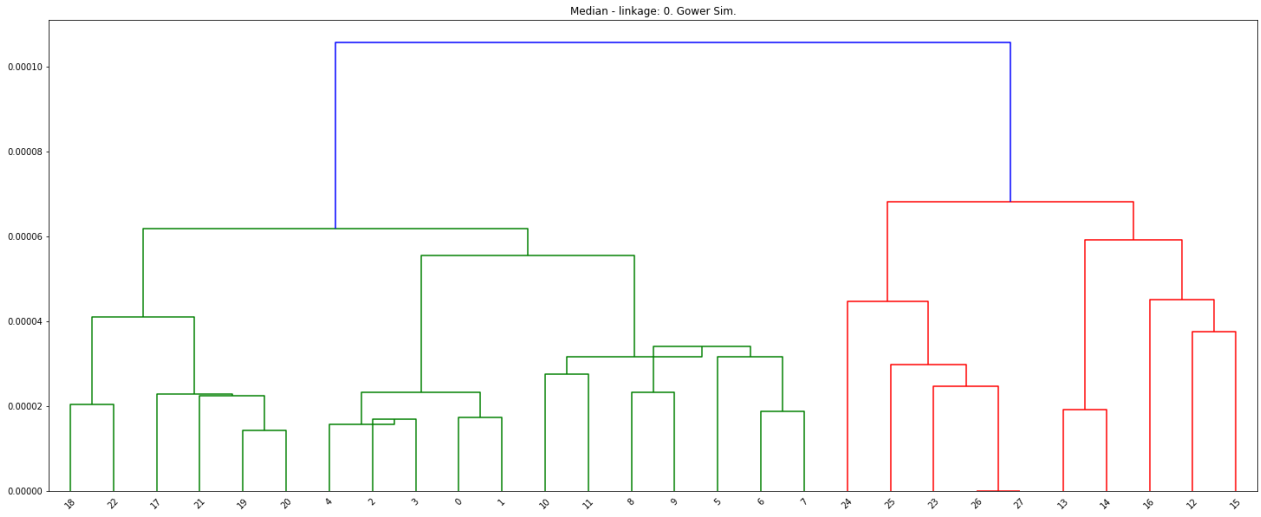
        ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim. Klaster trzeci rozbity na trzy mniejsze grupy. Ostatnie dwie grupy rozpoznane poprawnie.

c. Metoda pełnego wiązania:
    i. Dendrogram:



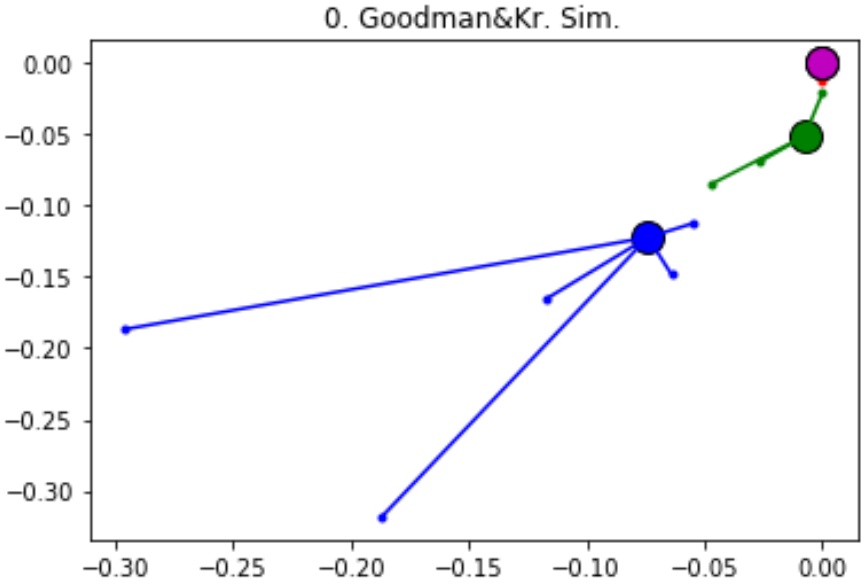
    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i czwartym. Klaster trzeci złączony z klastrem piątym.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Pierwszy klaster złączony z klastrem drugim i czwartym. Klaster trzeci złączony z klastrem piątym.

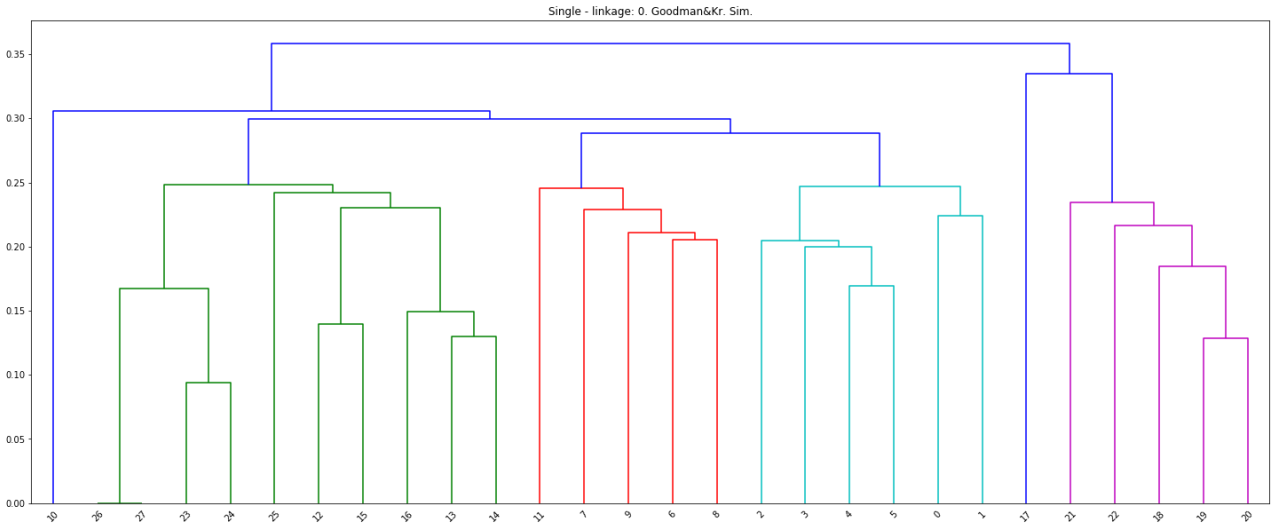
14. Współczynnik podobieństwa Goodmana & Kruskala:
    a. Propagacja powinowactwa:
        i. Wizualizacja:




```
Result: [0 0 0 0 0 0 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 2 4 4]
```

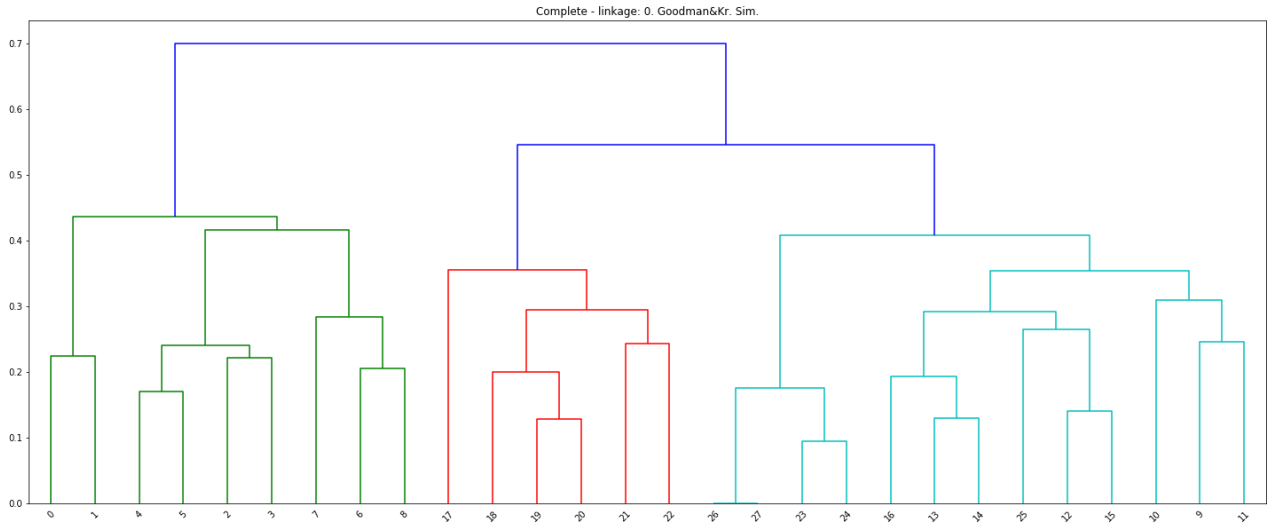
        ii. Poprawność: Liczba klastrów rozpoznana poprawnie. Złe przyporządkowanie elementów do grup.

    b. Metoda pojedynczego wiązania:
        i. Dendrogram:



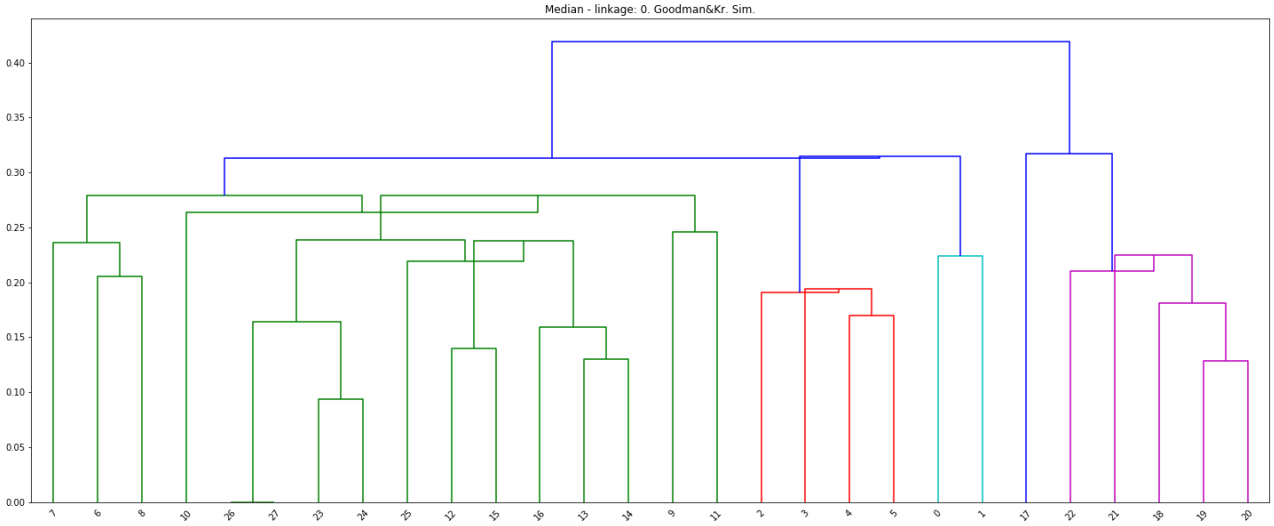
        ii. Poprawność: Źle rozpoznana ilość klastrów. Złe przyporządkowanie elementów do grup, dwa elementy nie przyporządkowane do żadnego ze skupień.

c. Metoda pełnego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Złe przyporządkowanie elementów do grup.

d. Metoda środkowego wiązania:
    i. Dendrogram:



    ii. Poprawność: Źle rozpoznana ilość klastrów. Złe przyporządkowanie elementów do grup.

Poniżej znajduje się podsumowanie wyników dla zbioru małych profili.

	AP	Single	Complete	Median
Odległość Hamminga	Tak	Nie	Nie	Nie
Odległość Vari'ego	Tak	Nie	Nie	Nie
Współczynnik różnicy Rogera – Tanimoto	Tak	Nie	Nie	Tak
Odległość różnicy rozmiaru	Tak	Nie	Nie	Nie
Odległość różnicy wzorca	Tak	Nie	Nie	Nie
Współczynnik różnicy Jaccarta – Needhama	Nie	Nie	Nie	Nie
Współczynnik różnicy Sokala – Sneatha	Nie	Nie	Nie	Nie
Współczynnik różnicy Russera – Rao	Nie	Nie	Nie	Nie
Współczynnik podobieństwa Sokala – Sneatha (IV)	Nie	Nie	Nie	Nie
Współczynnik podobieństwa Sokala – Sneatha (V)	Nie	Nie	Nie	Nie
Współczynnik różnicy Yule'a	Nie	Nie	Nie	Nie
Odległość Hellingera	Nie	Nie	Nie	Nie
Współczynnik podobieństwa Gowera	Nie	Nie	Nie	Nie
Współczynnik podobieństwa Goodmana & Kruskala	Nie	Nie	Nie	Nie

Tab. 1. Wyniki zrealizowanej analizy skupień dla zbioru małych profili.

Identyczna analiza skupień została przeprowadzona również dla zbioru dużych krzywych. W tym zbiorze do rozpoznania były dwie grupy. Poniżej znajduje się tabela przedstawiająca wyniki tej analizy:

	AP	Single	Complete	Median
Odległość Hamminga	Tak	Tak	Tak	Tak
Odległość Vari'ego	Tak	Tak	Tak	Tak
Współczynnik różnicy Rogera – Tanimoto	Tak	Tak	Tak	Tak
Odległość różnicy rozmiaru	Tak	Tak	Tak	Tak
Odległość różnicy wzorca	Tak	Tak	Tak	Tak
Współczynnik różnicy Jaccarta – Needhama	Tak	Tak	Tak	Nie
Współczynnik różnicy Sokala – Sneatha	Tak	Nie	Nie	Nie
Współczynnik różnicy Russera – Rao	Tak	Tak	Tak	Tak
Współczynnik podobieństwa Sokala – Sneatha (IV)	Tak	Tak	Tak	Tak
Współczynnik podobieństwa Sokala – Sneatha (V)	Tak	Tak	Tak	Tak
Współczynnik różnicy Yule'a	Tak	Tak	Tak	Tak
Odległość Hellingera	Tak	Nie	Nie	Nie
Współczynnik podobieństwa Gowera	Tak	Nie	Nie	Nie
Współczynnik podobieństwa Goodmana & Kruskala	Tak	Nie	Nie	Nie

Tab. 2. Wyniki przeprowadzonej analizy skupień dla zbioru dużych krzywych.

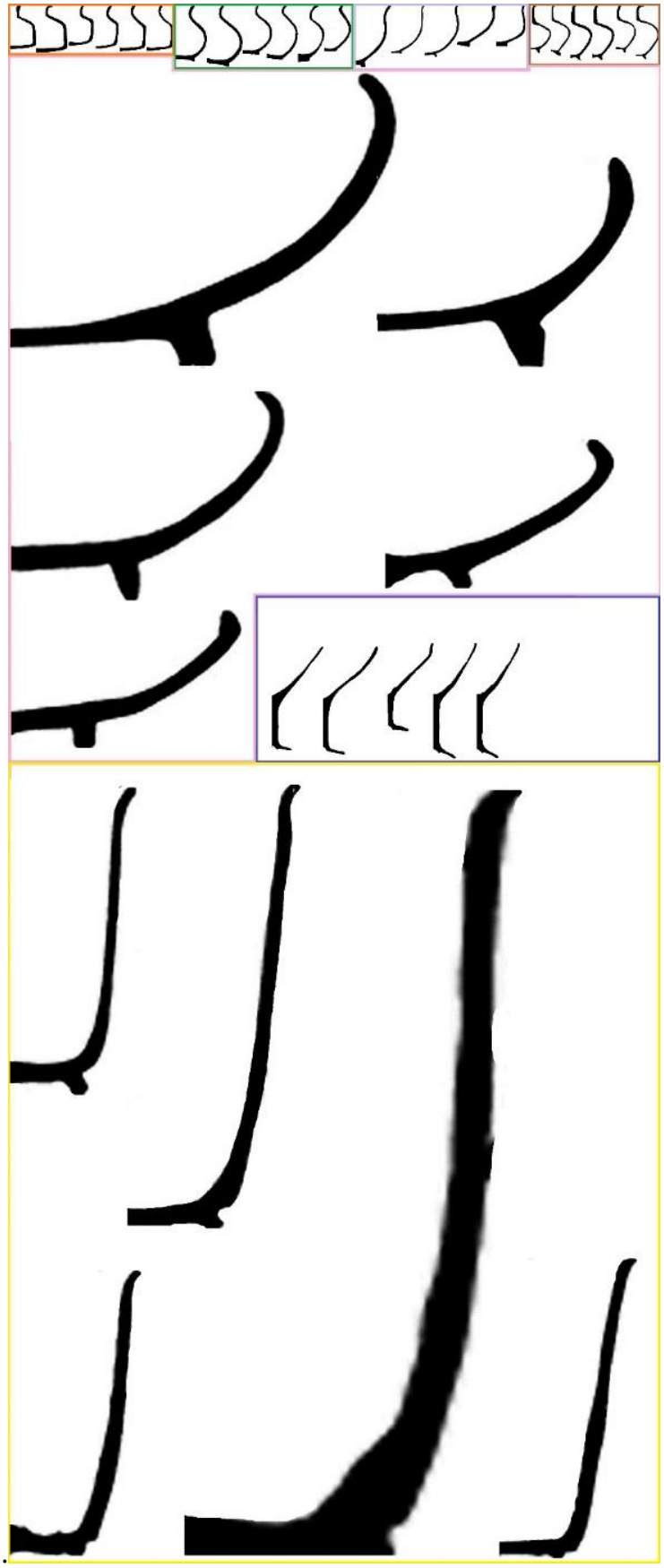
Początkowe rozróżnienie elementów na zbiory, które pozwoliło nie stracić informacji o rozmiarze elementów, równocześnie pozwoliło wyciągnąć dodatkowe wnioski. Okazuje się, że jeśli chodzi o dobór metod klastrowania, najbardziej uniwersalną pasującą do profili

ceramicznych jest algorytm propagacji koligacji. Jest ona wyjątkowo dobra w sytuacji małej ilości grup, ale dobrze radzi sobie w każdej ilości klastrów.

Jeśli chodzi o miary podobieństwa, najlepszy wynik został otrzymany przy zastosowaniu współczynnika różnicy Rogera – Tanimoto do analizy. Prawie tak samo dobry wynik otrzymany został przy użyciu odległości Vari'ego. Oznacza to, że ważnym jest, aby podczas analizy korzystać z pełnego spectrum informacji. Współczynniki różnicy Jaccarta – Needhama oraz Sokala – Sneatha nie biorą pod uwagę informacji o ilości nałożonych na siebie białych pikseli na obu obrazkach i, jak widać po wynikach analizy, jest to utrata zbyt cennej informacji. Odległość różnicy rozmiaru i różnicy wzorca to miary badające kwadraty informacji – na rozważanym zbiorze danych pochodzących z profili ceramicznych zastosowanie tej miary doprowadziło do otrzymania prawidłowego wyniku. Pozostałe miary, działające na niepełnej informacji bądź na jej iloczynie, pierwiastkach czy maksimach nie doprowadziły do poprawnego wyniku.

Kod źródłowy programu realizującego analizę skupień jest umieszczony pod linkiem: https://github.com/CzubajPaulina/pracamagisterska. Znajduje się on także w załączniku dołączonym do tej pracy. Opracowane narzędzie jest gotowe do uruchomienia i użycia do celów prywatnych.

Na następnej stronie znajduje się wizualizacja poprawnych wyników.



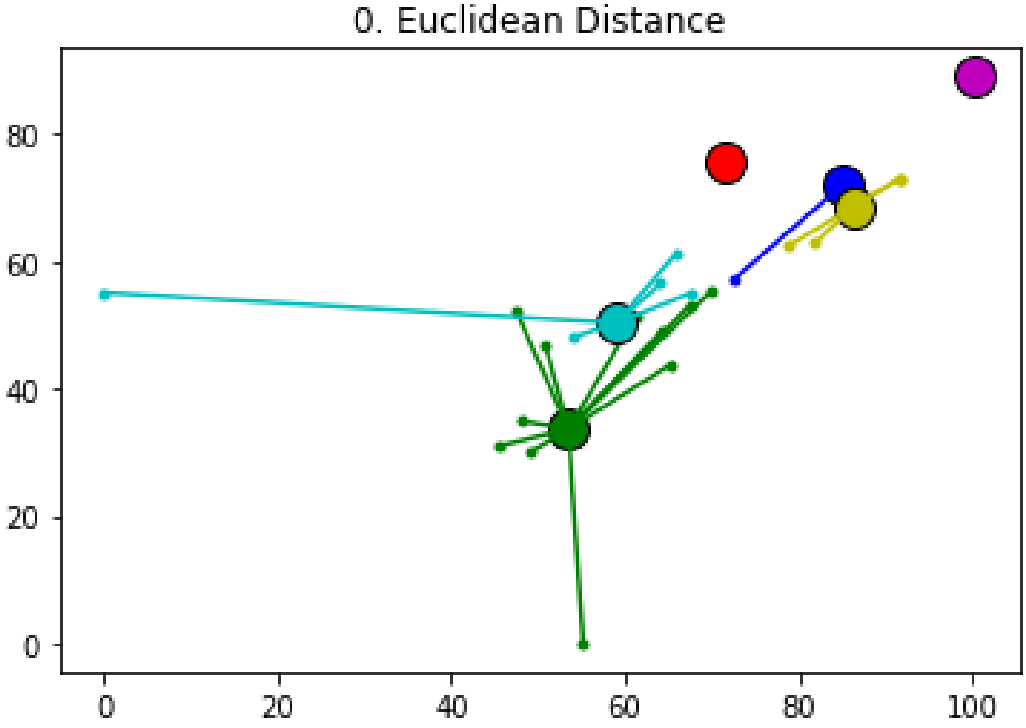
Rys. 8. Zbiór testowy pogrupowany.

Rozdział 5.
Analiza syntaktyczna

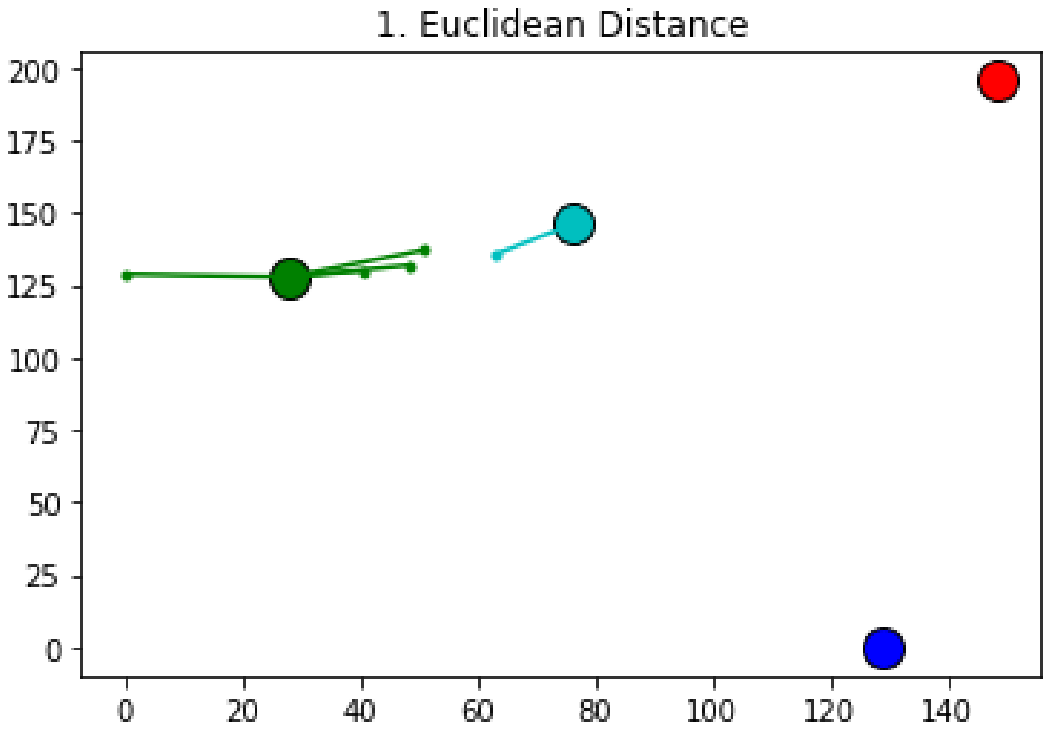
W tym rozdziale zostanie sprawdzone podejście analizy syntaktycznej do badanego zbioru profili ceramicznych. W tym celu na danych została przeprowadzona standardowa ich obróbka, na którą składa się: podział na grupy pod kątem wielkości, normalizacja, konwersja na macierze binarne. Następnie z profili została wydzielona oś grubości jednego piksela prowadząca przez środek krzywizny. Następnie, wygenerowanie gramatyki, reprezentowanej jako kody łańcuchowe Freemana, zostało, wg autorskiego pomysłu, opisane metryką euklidesową – profil został podzielony na 50 przedziałów i opisany za pomocą 50 liczb, każda z nich odwzorowuje kąt przejścia pomiędzy początkiem a końcem nowego przedziału. Następnie, słowa zostały pogrupowane za pomocą algorytmu propagacji powinowactwa.

Całość eksperymentu jest wykonana poprzez stworzenie kodu w języku programowania Python, wersja 3.6, kod uruchamiany był przy użyciu aplikacji/interpretera Jupyter Notebook.

Poniżej znajdują się wyniki procesu grupowania wyrazów pochodzących ze zbioru małych i dużych krzywych:



Wynik: `[3 1 0 1 1 1 0 0 1 1 1 2 1 1 1 1 1 3 3 3 3 4 3 5 5 5 5 5]`



Wynik: `[1 0 1 1 1 3 2 1 1 3]`

Okazuje się, że metoda generowania kodów łańcuchowych Freemana nie jest rozwiązaniem prowadzącym do otrzymania prawidłowych rezultatów na zbiorze profili ceramicznych. Powodem tego jest niedokładność w tworzeniu rysunków krzywizn oraz różnice w ich grubości. Z tego też powodu już wyciągnięcie profilu o grubości jednego piksela jest procesem narażonym na błędy prowadzące do wygenerowania błędnej osi. Wnioski te wynikają z przeprowadzonego eksperymentu.

Kod źródłowy programu realizującego analizę syntaktyczną jest umieszczony pod linkiem: https://github.com/CzubajPaulina/pracamagisterska. Znajduje się on także w załączniku dołączonym do tej pracy. Opracowane narzędzie jest gotowe do uruchomienia i przetestowania w celach prywatnych.

Rozdział 6.
Inne metody

Przy obecnym stanie rozwoju nauki dostępnych jest wiele metod analizy obrazu i widzenia komputerowego. W ostatnich czasach dynamicznie rozwijanym podejściem jest m.in. korzystanie z rozwiązań sztucznej inteligencji – konwolucyjnych sieci neuronowych. Jednak rozważany w tej pracy zbiór nie jest dedykowany do tego typu rozwiązań – składa się ze zbyt małej ilości elementów. Dostępność zbiorów jest niewielka, a te, które zostały udostępnione i z których korzystają naukowcy są małoliczne – rzędu 200 obrazków. Taka ilość blokuje użycie sztucznej inteligencji. Proces powielania danych poprzez ich przekształcanie – tzw. augmentacja, również nie doprowadziłaby do otrzymania poprawnych wyników. Z powodu małej ilości grup, powielenie profili mogłoby jedynie doprowadzić do przeuczenia modelu. Otrzymany model również nie byłby elastyczny do rozpoznawania nowych obrazków – podejścia nienadzorowanego.

Podsumowanie

Celem rozważanej pracy magisterskiej był dobór metod i przeprowadzenie próby pogrupowania zbioru krzywych reprezentujących profile naczyń ceramicznych, co dążyło do stworzenia narzędzia do automatycznego klastrowania. Dokonano przeglądu literatury i zdefiniowania problemu. Wybrane metody zostały zewaluowane pod kątem ich dedykowania do rozważanych danych. Przeprowadzono proces wstępnej obróbki danych. Łącząc tradycyjne metody i podejście autorskie dokonano analizy skupień i analizy syntaktycznej zbioru. Otrzymane wyniki są wynikami poprawnymi, co potwierdza słuszność przeprowadzonych prac.

Przeprowadzona analiza problemu nie wyczerpuje zbioru możliwych jego rozwiązań. Kierunkiem badawczym może być próba wykorzystania topologicznej analizy danych czy też prace nad rozbudowaniem zbioru tak, aby możliwe stało się zastosowanie do problemu algorytmów sztucznej inteligencji.

Bibliografia

1. A. Karasik, U. Smilansky, *3D scanning technology as a standard archaeological tool for pottery analysis: practice and theory*, Journal of Archaeological Science 35(5), 2008.
2. L. Drabik, A. Kubiak-Sokół, E. Sobol, *Słownik języka polskiego PWN,* Wydawnictwo Naukowe PWN, 2017.
3. U. Schurmans, A. Razdan, A. Simon, M. Marzke, P. McCartney, *Advances in geometric modeling and feature extraction on pots, rocks and bones for representation and query via the Internet*, Computer Applications in Archaeology (CAA), 2001.
4. A. Razdan, D. Liu, M. Bae, M. Zhu, G. Farin, *Using geometric modeling for archiving and searching 3D archaeological vessels*, International Conference on Imaging Science, Systems, and Technology CISST, 2001.
5. F. Leymarie, D. Cooper, M. Joukowsky, B. Kimia, D. Laidlaw, D. Mumford, E. Vote. *The SHAPE Lab: New technology and software for archaeologists*, Computer Applications and Quantitative Methods in Archaeology, 2000.
6. K. Adler, M. Kampel, R. Kastler, M. Penz, R. Sablatnig, K. Schindler, S. Tosovic, *Computer Aided Classification of Ceramics: Achievements and Problems*, Sixth International Workshop on Archaeology and Computers, 2002.
7. H. Mara, R. Sablatnig, A. Karasik, U. Smilansky, *The uniformity of wheel produced pottery deduced from 3D image processing and scanning,* Imaging in Media and Education, 28th Workshop of the Austrian Association for Pattern Recognition, 2004.
8. H. Mara, R. Sablatnig, *Semiautomatic and automatic profile generation for archaeological fragments,* Proceedings of 10th Computer Vision Winter, 2005.
9. A. Gilboa, I. Sharon, *An archaeological contribution to the early Iron Age chronological debate: alternative chronologies for Phoenicia and their effects on the Levant, Cyprus and Greece,* Bulletin of the American Schools of Oriental Research, 2003.
10. A. Gilboa, A. Karasik, I. Sharon, U. Smilansky, *Towards computerized typology and classification of ceramics,* Journal of Archaeological Science 31, 2004.
11. P. Aparajeya, C. Piccoli, G. Papadopoulos, *Towards the automatic classification of pottery sherds: two complementary approaches,* Computer Applications and Quantative Methods in Archaeology Conference, 2013.
12. K. Konopka, D. Riegert, D. Kobylińska, U. Kobyliński, *Finds of pottery from archeological sites as a subject of materials science,* Inżynieria Materiałowa vol. 32, 2011.

13. H. Stoksik, *Technologia warsztatu ceramicznego średniowiecznego Śląska w świetle badań specjalistycznych i eksperymentalnych*, Wydawnictwo PWT we Wrocławiu, 2007.
14. M. Makridis, P. Daras, *Automatic Classification of Archaeological Pottery Sherds*, ACM Journal on Computing and Cultural Heritage 5, no. 4, 2012.
15. A. Karasik, U. Smilansky. *Computerized Morphological Classification of Ceramics*, Journal of Archaeological Science 38, no. 10, 2011
16. A. Martinez-Carrillo, *Computer Applications in Archaeological Pottery: a Review and New Perspectives*, On the Road to Reconstructing the Past, 36th Computer Applications and Quantitative Methods in Archaeology Conference, 2011
17. C. Piccoli, P. Aparajeya, G. Papadopoulos, J. Bintli, F. Leymarie, P. Bes, *Towards the automatic classication of pottery sherds: two complementary approaches*, Across Space and Time, 41st Conference on Computer Applications and Quantitative Methods in Archaeology, 2015.
18. C. Maiza, V. Gaildart, *Automatic classification of archaeological potsherds*, The Eight International Conference on Computer Graphics and Artificial Intelligence, 2005.
19. P. Mountjoy, *Regional Mycenaean Decorated Pottery*, Annual of the British School at Athens, 1990.
20. P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, 2005.
21. S. Choi, S. Cha, C. Tappert, *A Survey of Binary Similarity and Distance Measures*, Journal of Systemics, Cybernetics and Informatics, 2009.
22. B. Zhang, S. Srihari, *Properties of Binary Vector Dissimilarity Measures*, 2003.
23. Numpy and Scipy Documentation, https://docs.scipy.org/doc/ [dostęp: 27.05.2019].
24. K. Wang, J. Zhang, D. Li, X. Zhang, *Adaptive Affinity Propagation Clustering*, Acta Automatica Sinica 33(12), 2008.
25. B. Frey, D. Dueck, *Clustering by Passing Messages Between Data Points*, Science, 2007.
26. B. Everitt, S. Landau, M. Leese, *Cluster Analysis*, 2001.
27. L. Rutkowski, *Metody i techniki sztucznej inteligencji*, PWN, 2005.
28. S. Wierzchoń, M. Kłopotek, *Algorithms of Cluster Analysis*, PAN, 2015.
29. M. Flasiński, *Wstęp do sztucznej inteligencji*, PWN, 2011.
30. R. Tadeusiewicz, *Rozpoznawanie obrazów*, PWN, 1991.
31. Scikit-learn Documentation, https://scikit-learn.org/stable/documentation.html# [dostęp: 27.05.2019].
32. T. Pavlidis, *Algorithms for Shape Analysis of Contours and Waveforms*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980.

Załączniki

Do pracy dołączony jest skompresowany plik Praca_Magisterska_Czubaj_Załączniki, zawierający:

1. Pliki zawierające kody źródłowe programów:
    a. analiza_skupien.py,
    b. analiza_syntaktyczna.py.
2. Plik zawierający wywołanie kodów źródłowych: index.ipynb.
3. Plik zawierający dokładny opis zawartości skompresowanego pliku: read_me.txt.
4. Plik stanowiący dokumentację techniczną (zawierający m.in. opis instalacji obligatoryjnego oprogramowania): dokumentacja_techniczna.pdf.
5. Plik stanowiący dokumentację użytkową (zawierający opis procesu uruchamiania programów): dokumentacja_użytkowa.pdf.
6. Zbiór 38 profili wykorzystywany w pracy magisterskiej: dataset-original-size.rar.
7. Plik stanowiący opis zbioru testowego: opis_zbioru.pdf.