Bioinformatyka Sprawozdanie 5

Klastrowanie sekwencji. Algorytm UPGMA

Michał Marciniak 244811

1 Kod do repozytorium

`https://gitlab.com/MMarciniak103/bioinformatics`

2 Analiza złożoności obliczeniowej czasowej i pamięciowej:

Algorithm 1 Algorytm UPGMA

procedure INICJALIZACJA
    $X \leftarrow$ zbiór sekwencji
    $D \leftarrow$ macierz odległości pomiędzy sekwencjami (Utworzona np. przy wykorzystaniu algorytmu STAR)
    $C \leftarrow$ zbiór klastrów - początkowo każdy klaster to osobna sekwencja
    $d(x, y) \leftarrow$ funkcja wyznaczająca odległość pomiędzy sekwencjami x i y
procedure TWORZEDNIE DENDROGRAMU
    while $\exists x \in X \notin C$ **do**
        Znalezienie najbardziej podobnych do siebie klastrów:
        $a^*, b^* = \text{argmin}_{a,b} \frac{1}{|a| \cdot |b|} \sum_{x \in a} \sum_{y \in b} d(x, y)$
        utworzenie nowego klastra c:
        $c = a^* \cup b^*$
        $D \leftarrow$ zaktualizowana macierz kosztów, uwzględniająca nowy klaster
        $C+ = c$

Złożoność obliczeniowa i pamięciowa wynosi : $O(n^2) \leftarrow$ wykonujemy n-1 iteraci, a każda z nich ma złożoność O(n)

3 Testy akceptacyjne

Bardzo dobrym pomysłem jest testowanie algorytmu na danych słowach pisanych w różnych językach. Jest to bardziej intuicyjne dla osoby sprawdzającej, wobec czego można z góry określić spodziewany wynik. W związku z tym część testów została przeprowadzona właśnie w takiej formie.

Identyfikator	Cel testu	Wprowadzone dane	Spodziewany wynik
1	test algorytmu	pol: kawa ang: coffe ger: kaffe fr: cafe czh: kava	(fr,(pol,czh)),(ang,ger))
2	test algorytmu	pol: lampa ang: lamp ger: lampe fr: lampe czh: svitilna	(czh,(ang,(pol,(ger,fr))))
3	test algorytmu	pol: zupa ang: soup ger: souppe fr: soupe czh: polevka	(czh,(pol,(ger,(ang,fr))))
4	test algorytmu	x1: GCAGCA x2: GCCGCA x3: TACAA x4: AAACA x5: AAACA	((x1,x2),(x3,(x4,x5)))
5	test algorytmu	pol: banan ang: banana ger: banane fr: banane czh: banan	((ang,(ger,fr)),(pol,czh))
6	test algorytmu	pol: morze ang: sea ger: meer fr: mer czh: more	((pol,czh),(ger,(ang,fr)))

Figure 1: Testy akceptacyjne

Wszystkie testy wykonane zostały dla następujących kosztów:

- match - 0
- subsection - 1
- indent - 2

Ciekawym spostrzeżeniem jest to, że nie licząc pojedyńczych słów, język czeski najczęściej występuje jako klaster najbardziej oddalony od pozostałych klastrów.

3.1 Test 1

Figure 2: Wynik 1 testu

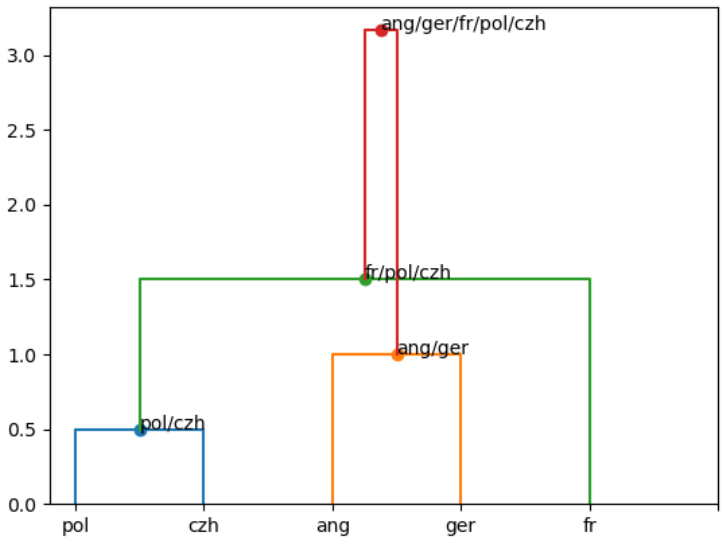

Figure 3: test 1 - dendrogram

3.2 Test 2

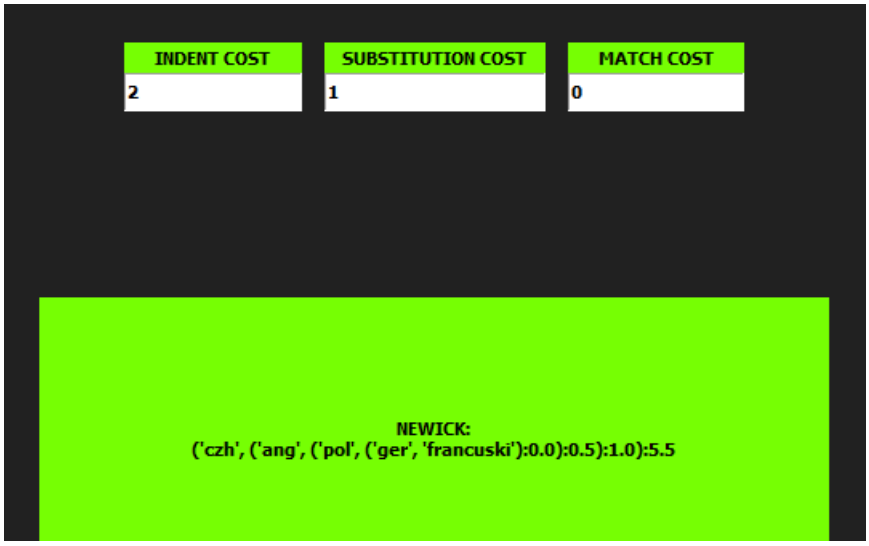

Figure 4: Wynik 2 testu

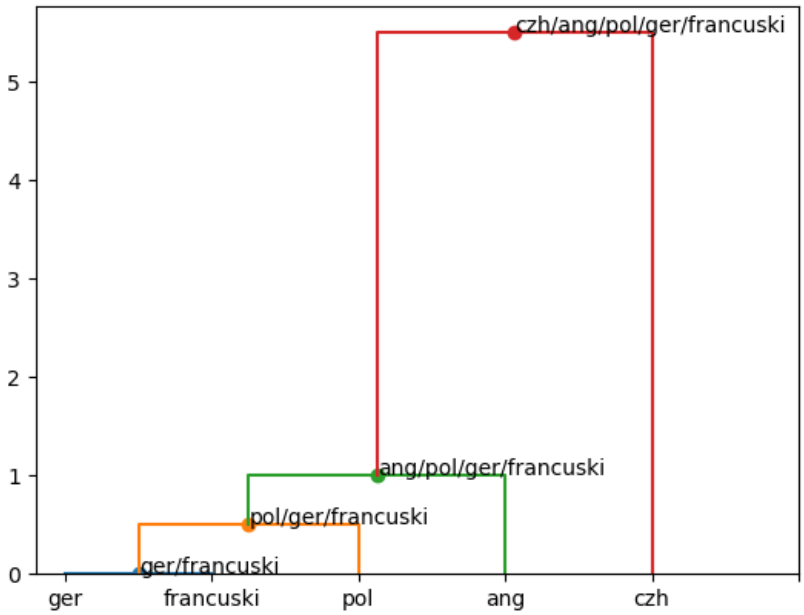

Figure 5: test 2 - dendrogram

3.3 Test 3

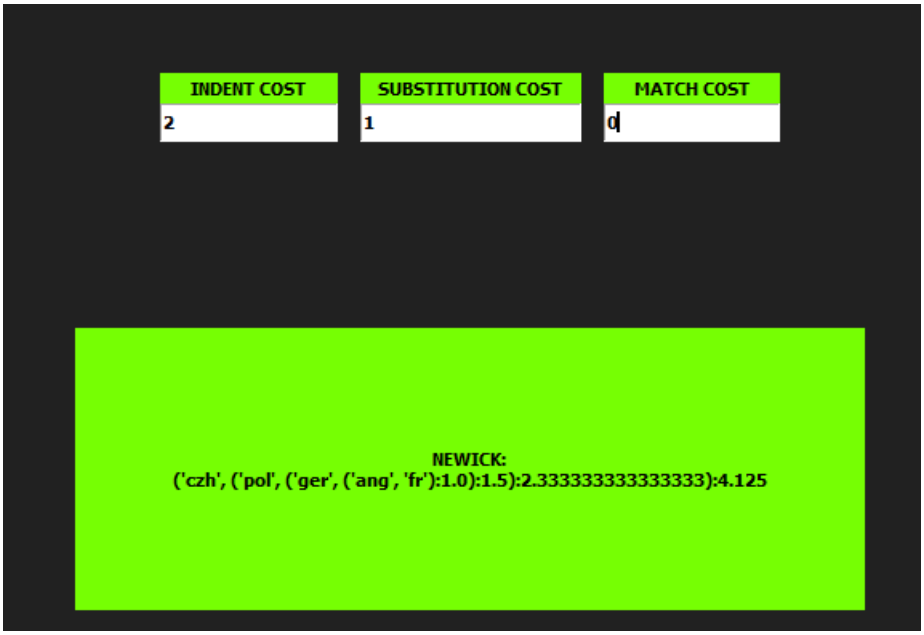

Figure 6: Wynik 3 testu

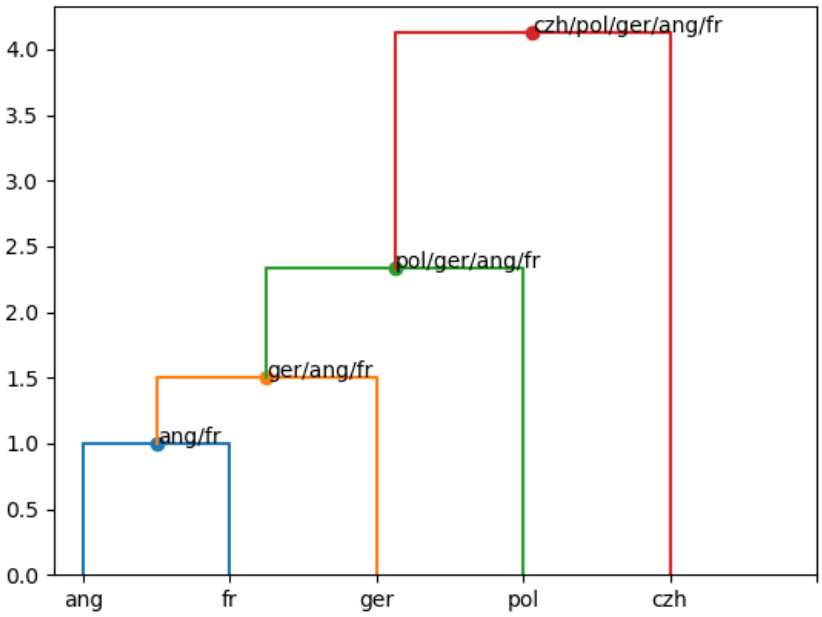

Figure 7: test 3 - dendrogram

3.4 Test 4

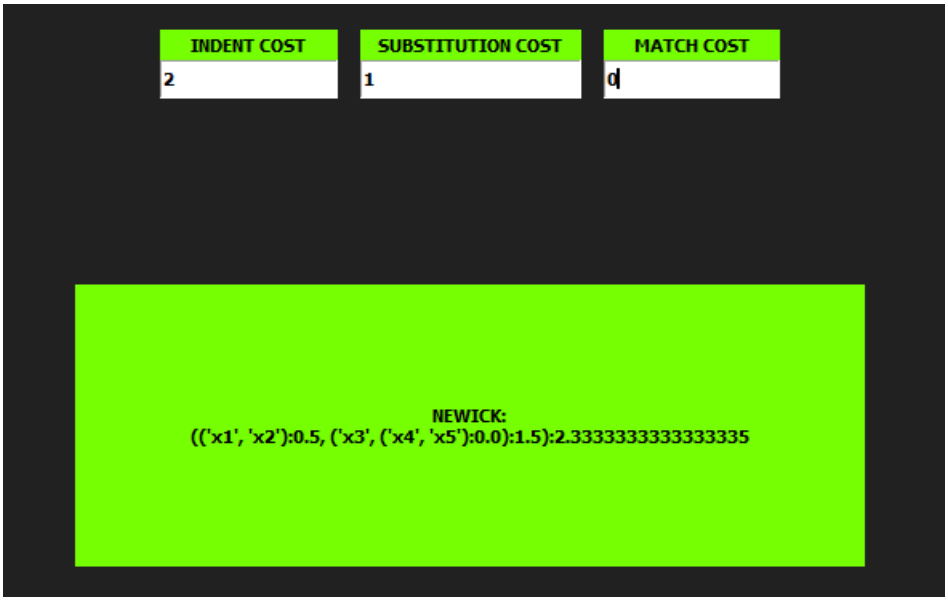

Figure 8: Wynik 4 testu

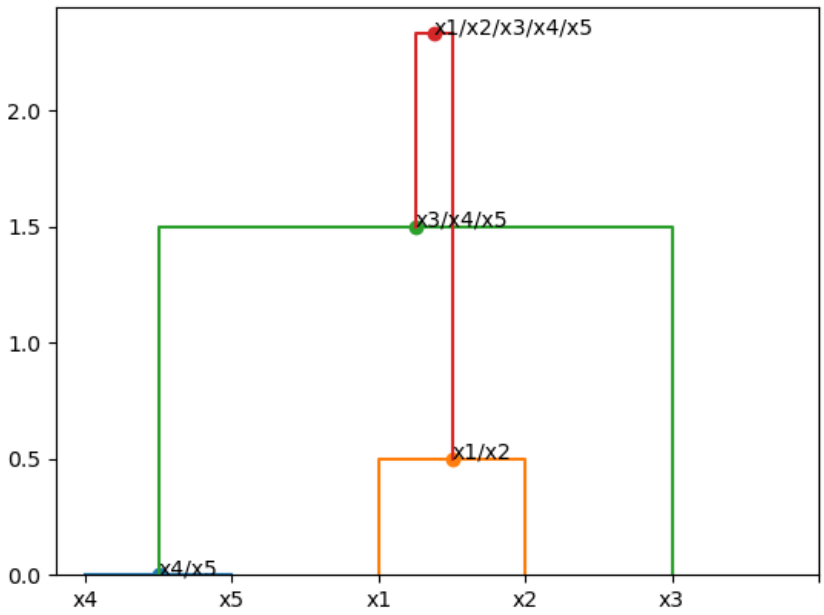

Figure 9: test 4 - dendrogram

3.5 Test 5

Figure 10: Wynik 5 testu

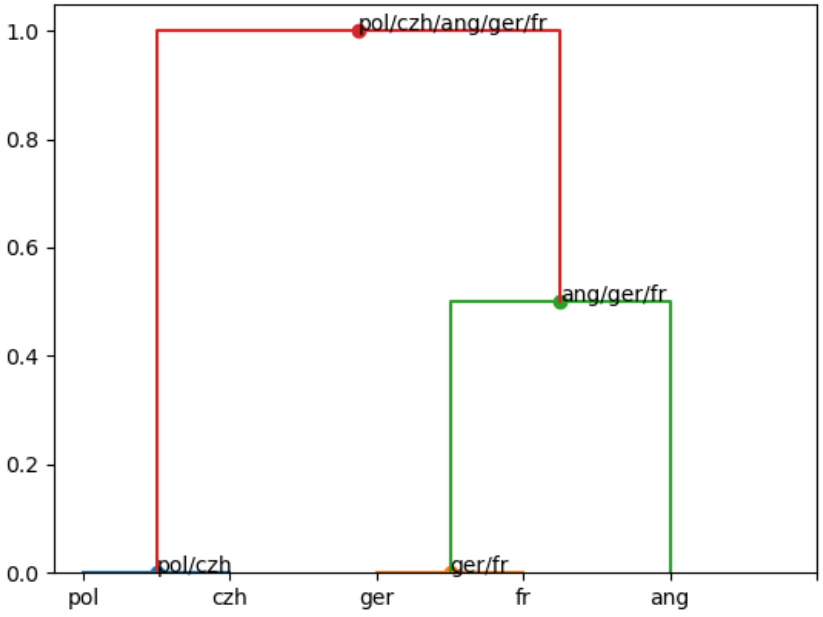

Figure 11: test 5 - dendrogram

3.6 Test 6

Figure 12: Wynik 6 testu

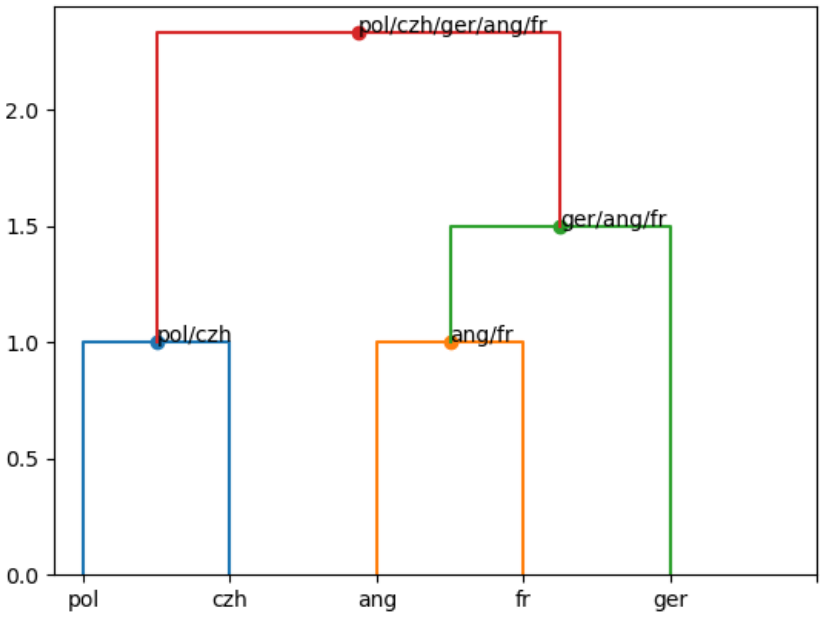

Figure 13: test 6 - dendrogram