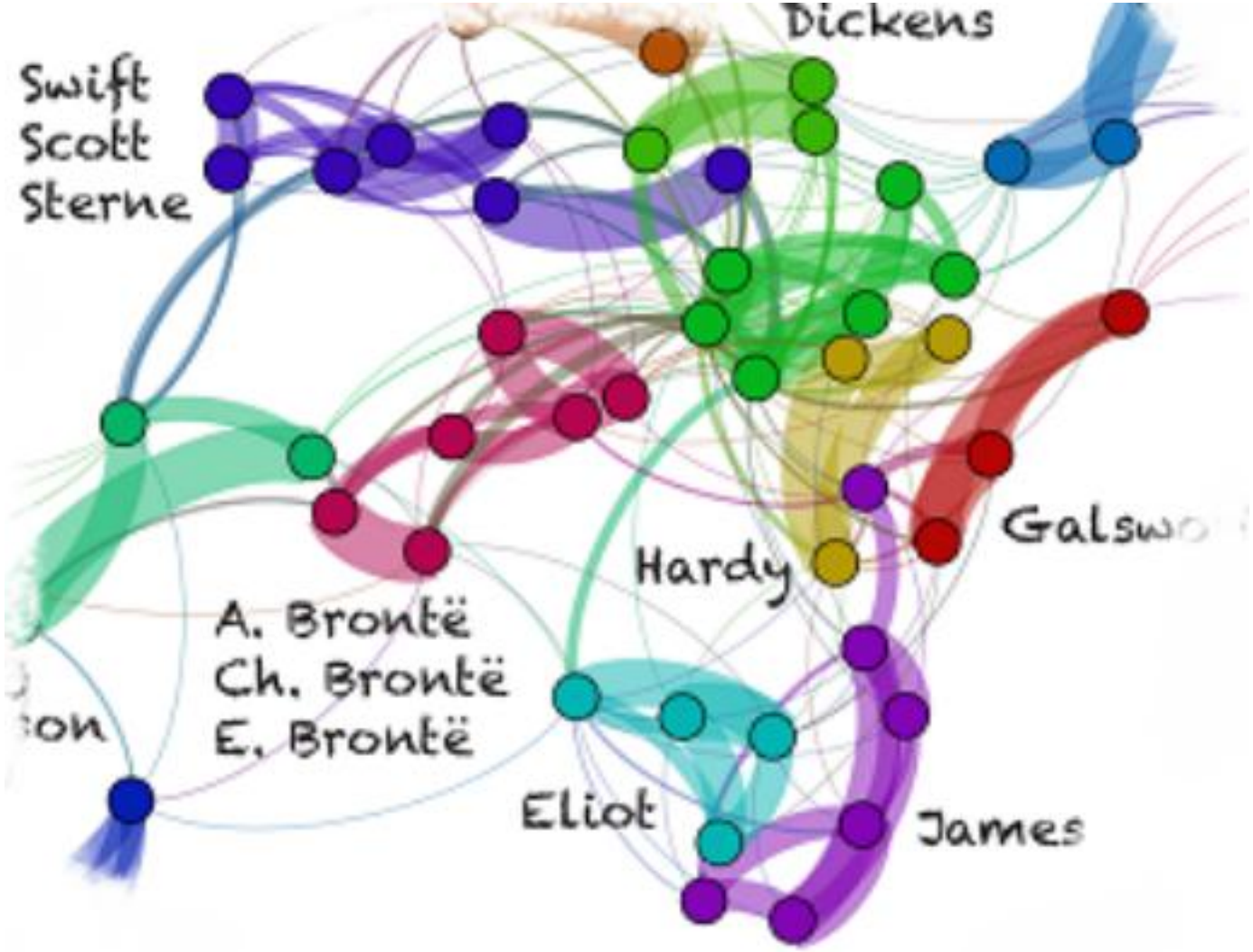
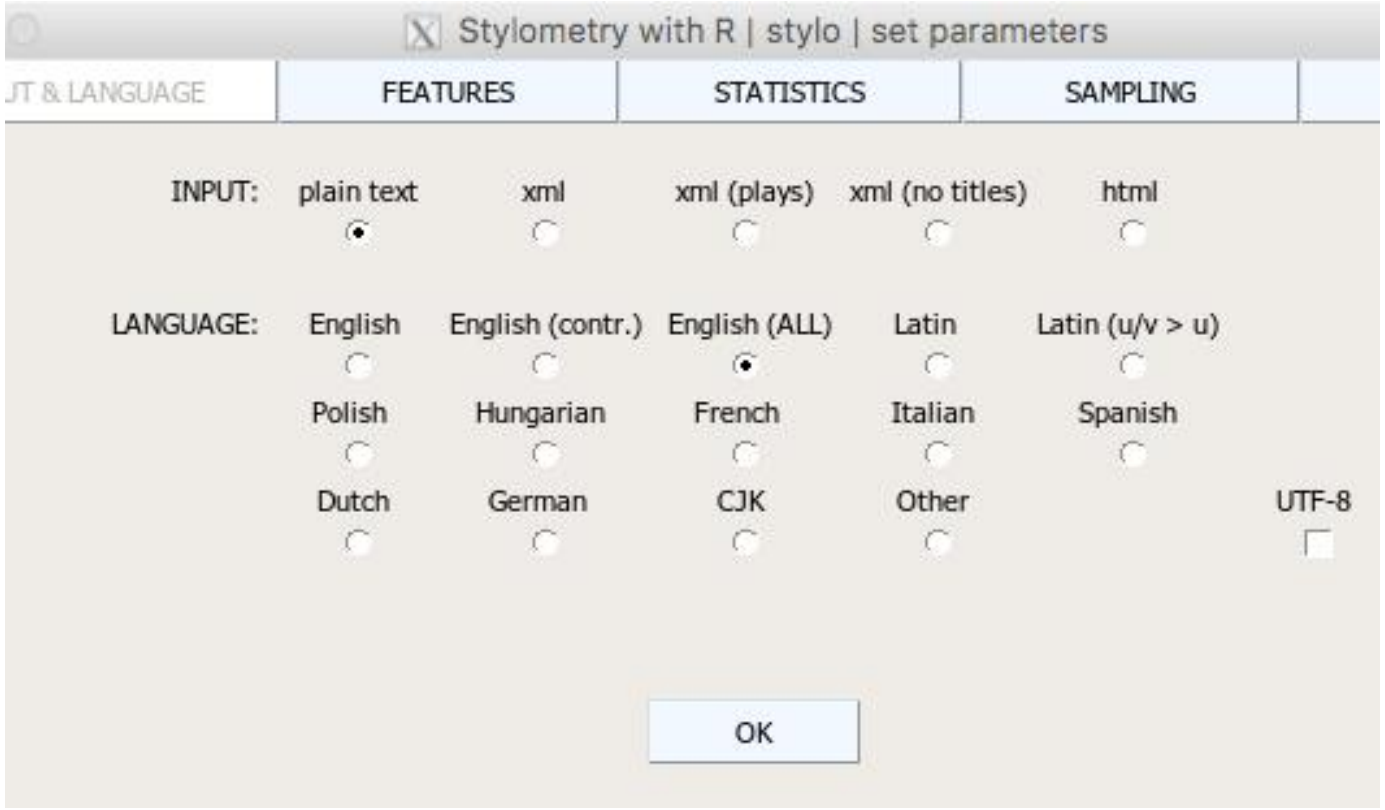
Stylometria z pakietem Stylo

Stylometria jest metodą analizy dzieł sztuki. Pozwala ona na określenie stylu danego dzieła, i na podstawie tego – kto jest jego potencjalnym twórcą, albo czy dana osoba naprawdę jest autorem. W ramach niniejszego projektu w celu wykonania takiej analizy skorzystano z pakietu „Stylo”.

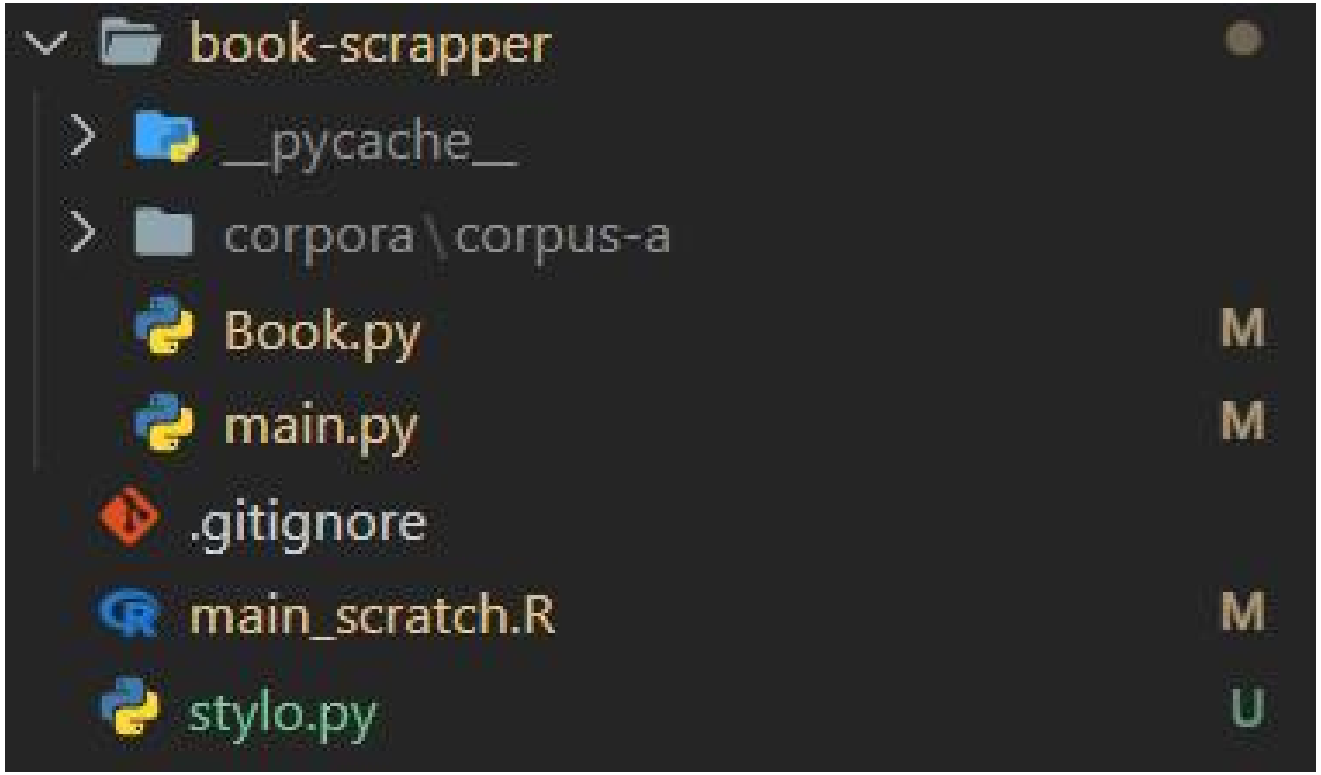


Czym jest Stylo

- Pakiet stylo został opracowany w 2016 przez Macieja Edera. Jest to potężne narzędzie do przeprowadzenia różnorodnych testów i analiz na zbiorach tekstowych. Wybór algorytmów i innych ustawień odbywa się przez interfejs graficzny, ale możliwe jest też ręcznie ustawianie parametrów w kodzie programu. Pakiet również przedstawia liczne metody wizualizacji wyników, co znacznie upraszcza ich odczyt i zrozumienie.

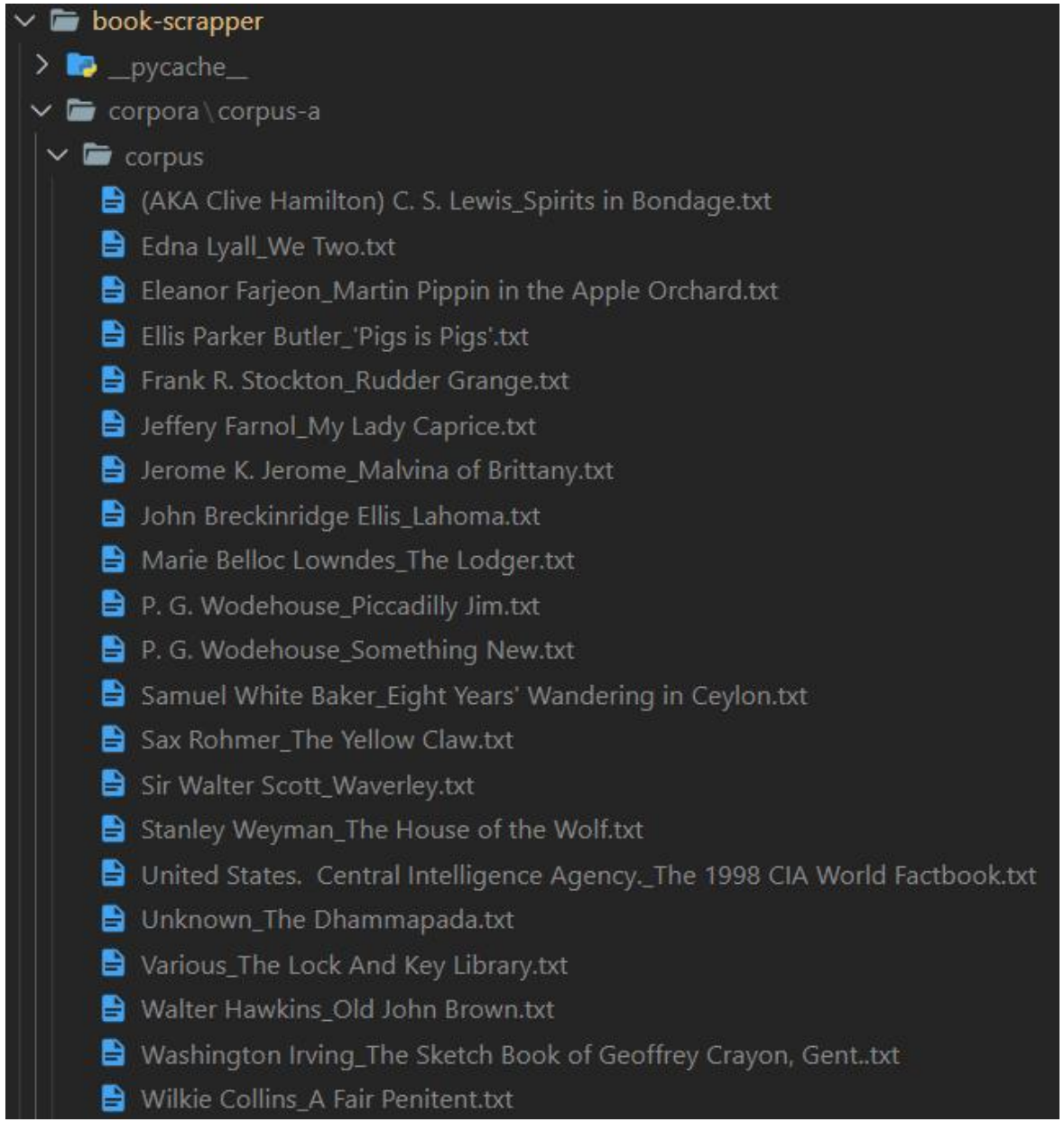


Praca ze Stylo



- Ważnym etapem pracy ze Stylo jest przygotowanie danych. Pakiet zawiera kilka przykładowych zbiorów tekstów i ich cech na potrzeby testowania, ale wraz z tym narzuca pewną „architekturę” organizacji projektu.

W naszym projekcie stworzyliśmy web scrapper, przy pomocy którego zostały pobrane teksty różnych autorów. Ograniczyliśmy się do 21 tekstu.



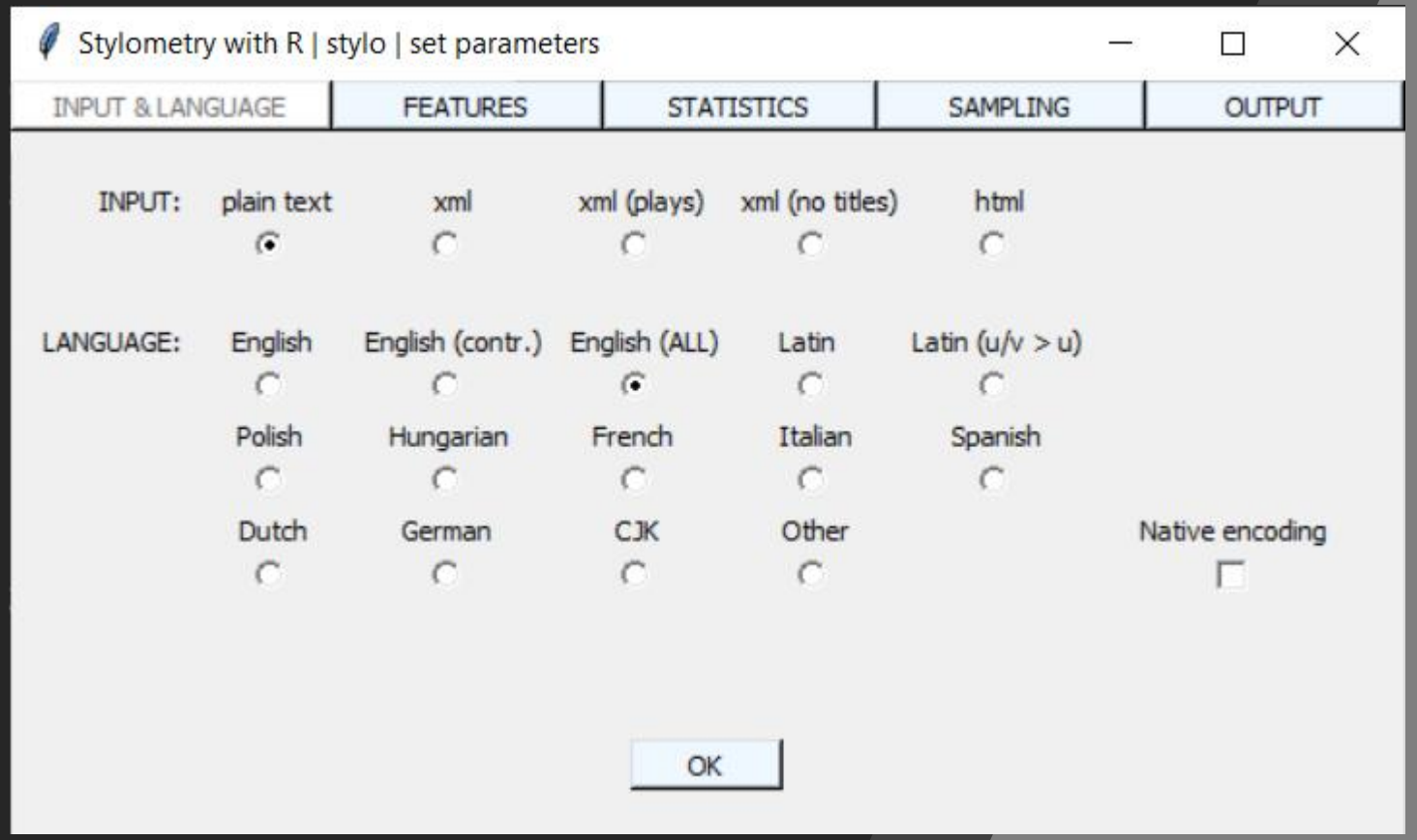
Następnym etapem jest uruchomienie Stylo albo z poziomu R, albo z poziomu Python.

```
library(stylo)
stylo(analysis.type = "CA",
write.png.file = TRUE, custom.graph.title = "CA Analysis",
gui = FALSE)
```

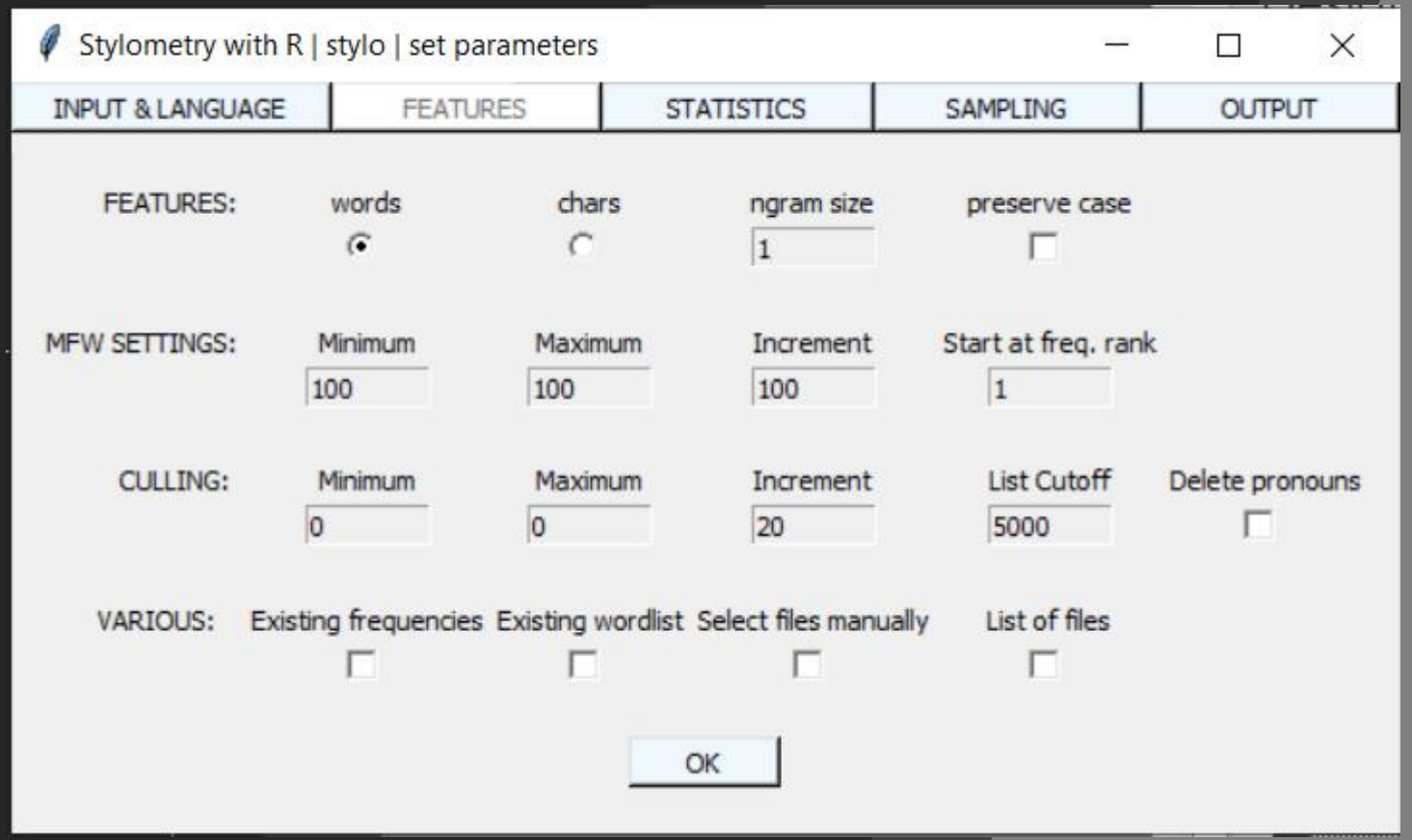
```
from rpy2 import robjects

R = robjects.r
R.library("stylo")
R.stylo(gui = True)
```

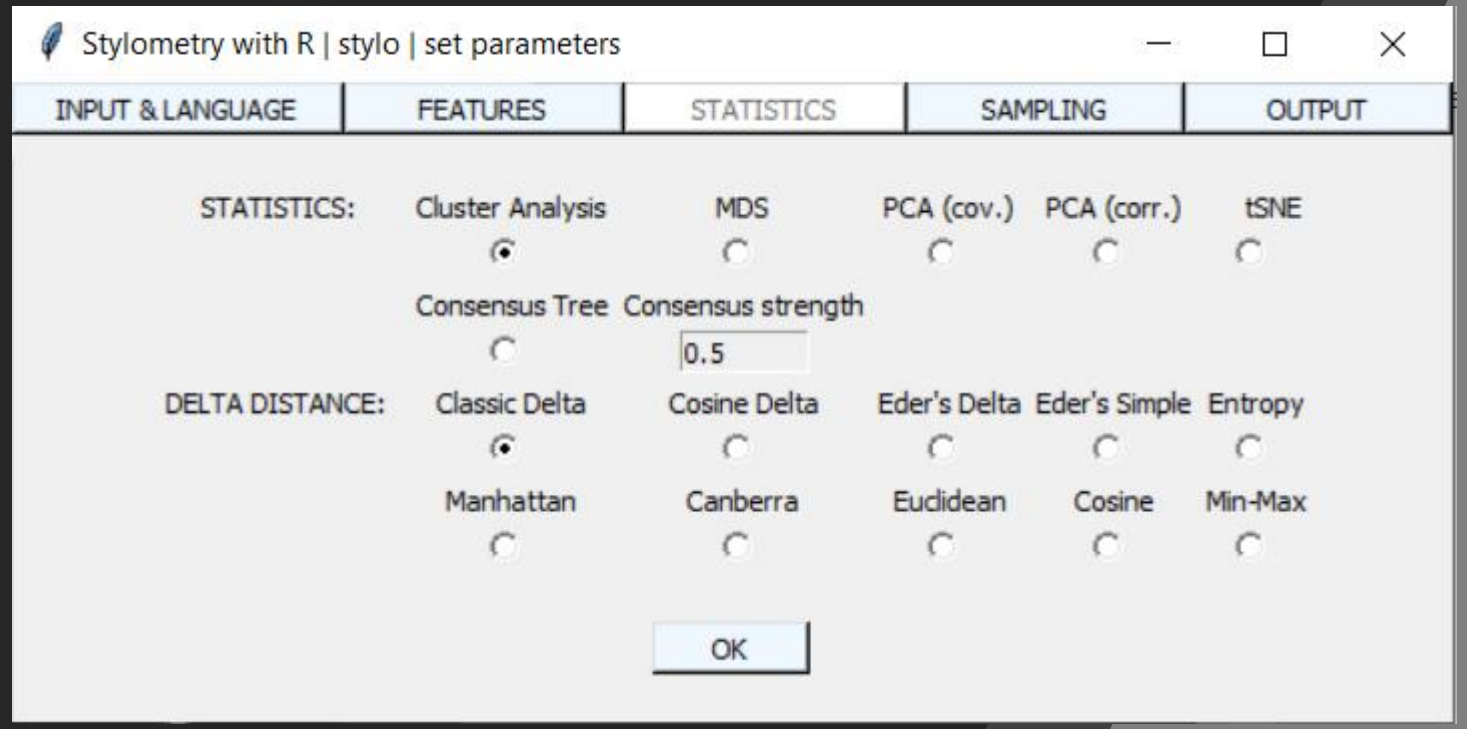
Ustawienia: język i dane wejściowe



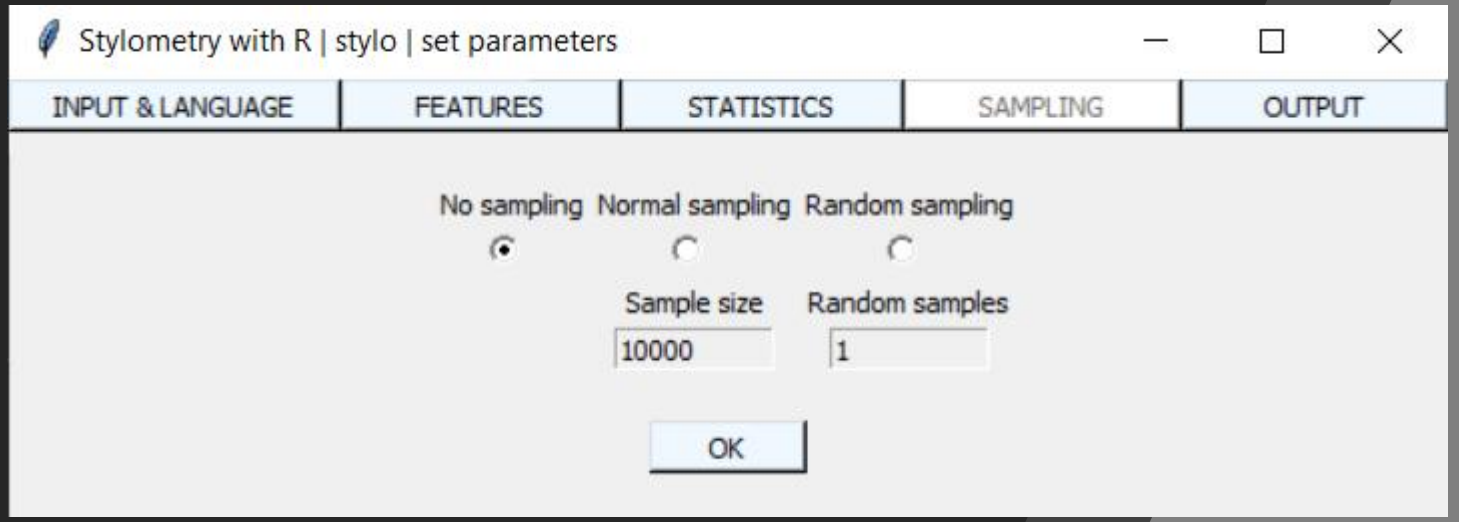
Ustawienia: cechy



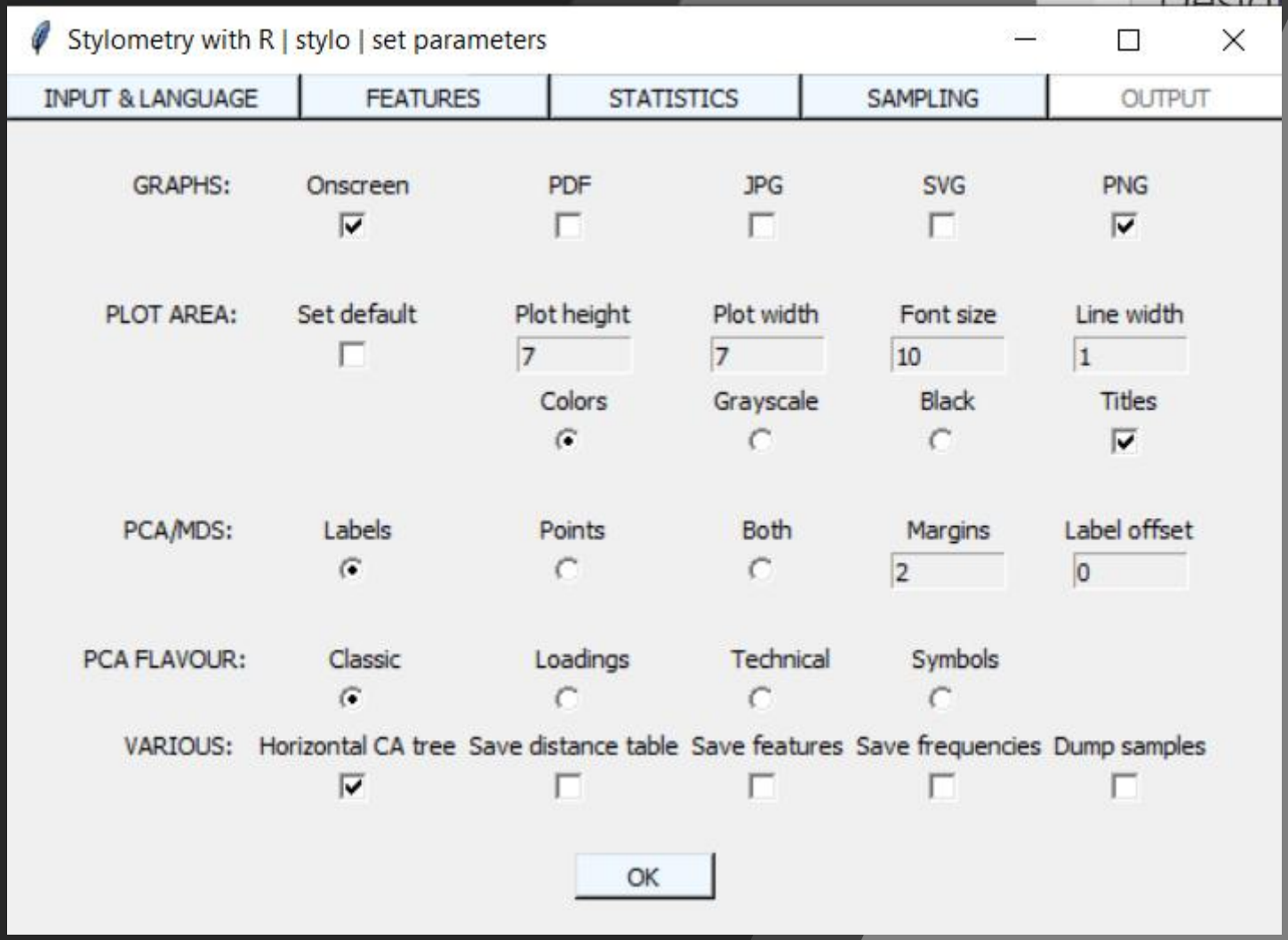
Ustawienia: statystyki i algorytmy



Ustawienia: pobieranie próbek



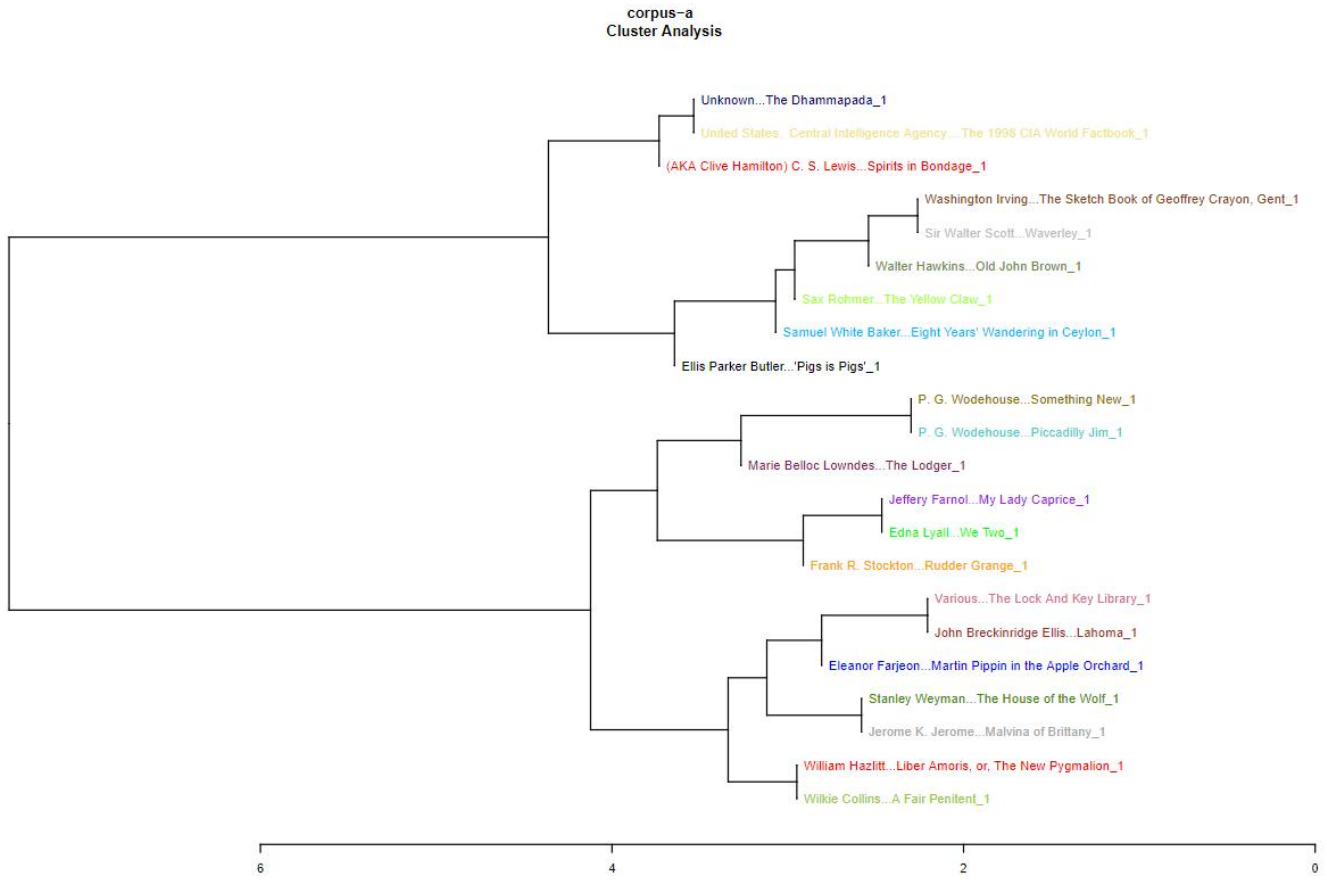
Ustawienia: zapis wyników



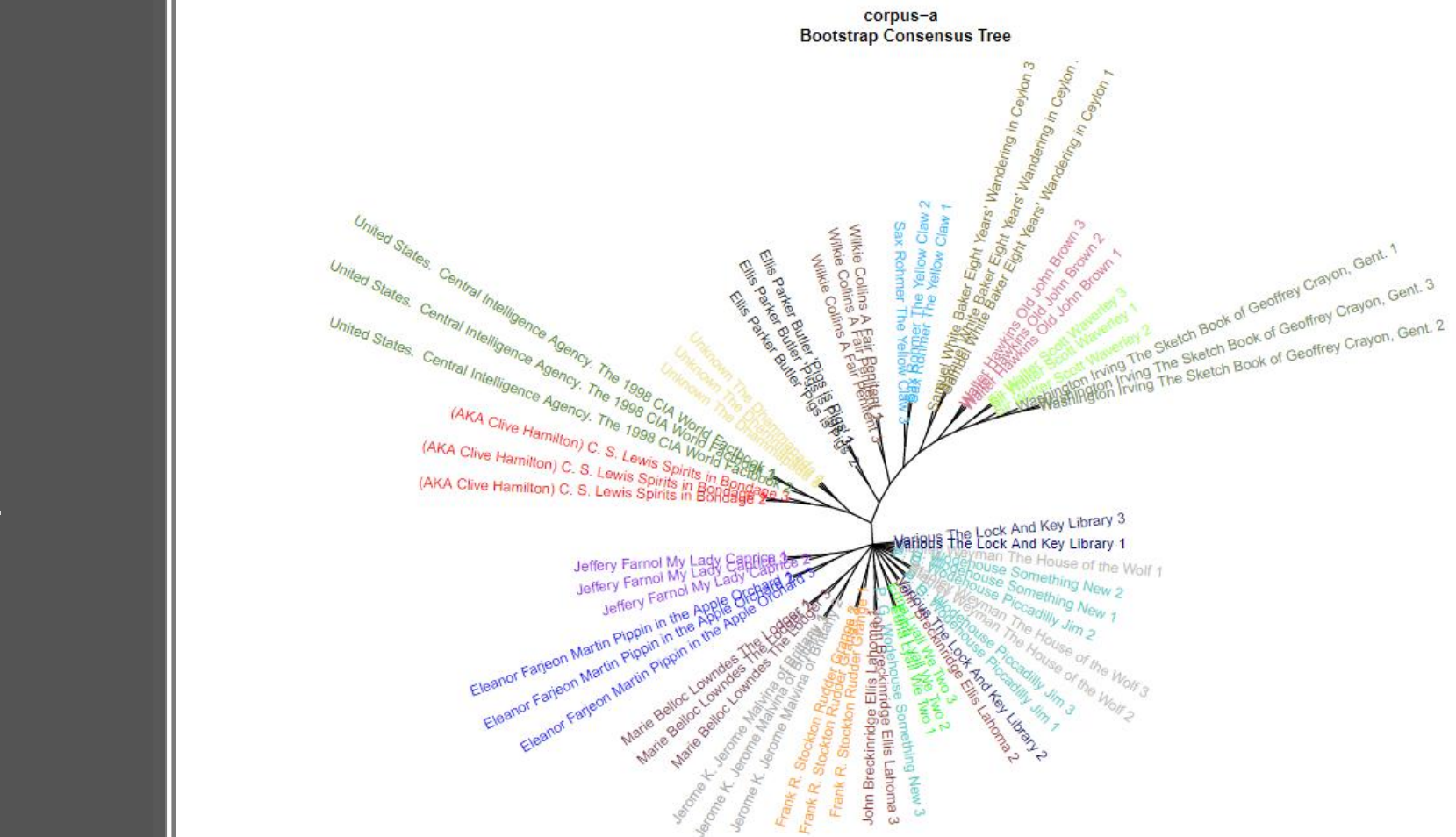
Co potrafi Stylo?

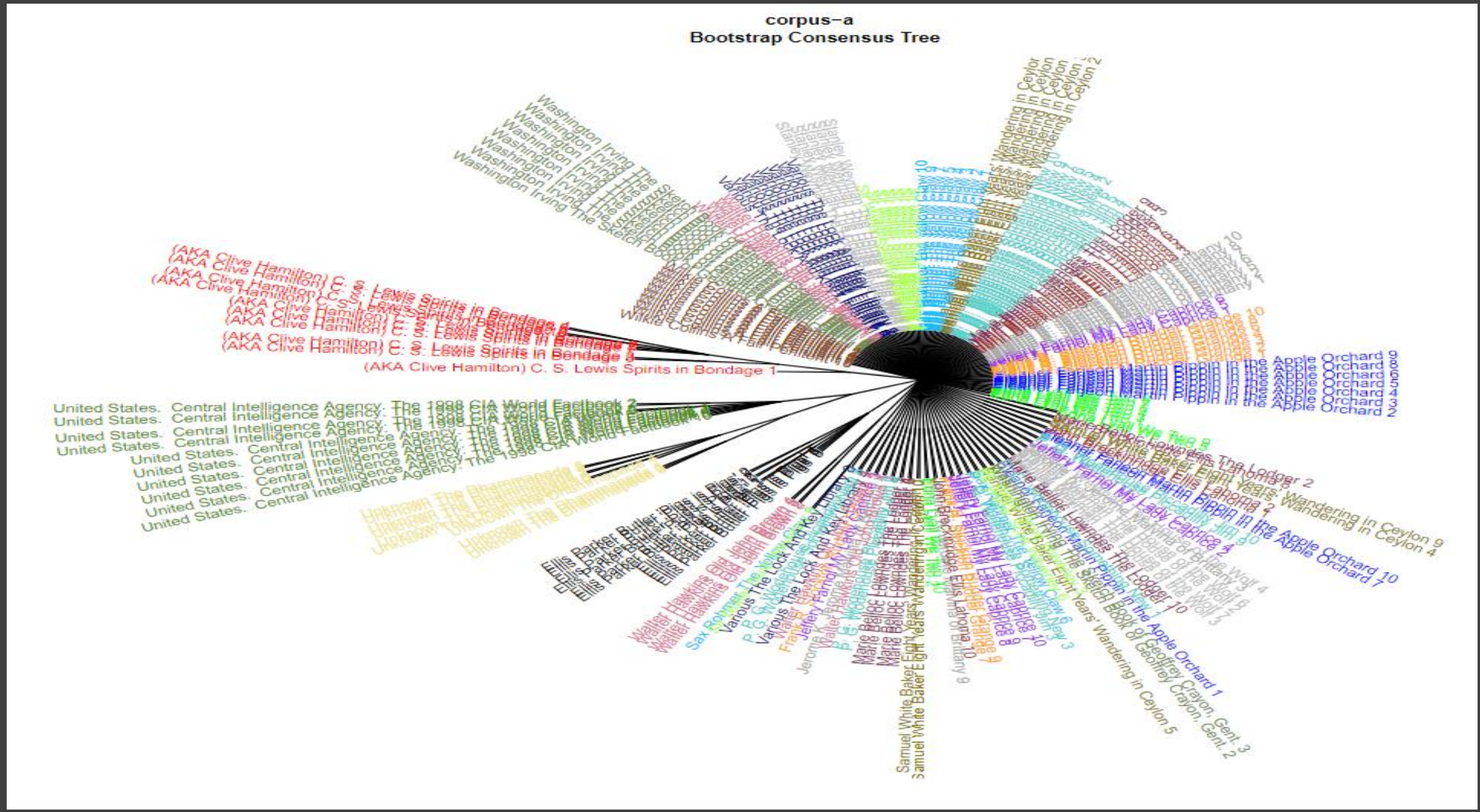
- Obliczanie podobieństw tekstów z wykorzystaniem różnych metryk (odległość kosinusowa, entropia …)
- Testowanie, kto jest autorem tekstu
- Wykrywanie plagiatu na podstawie analizy segmentów tekstu i uczenia maszynowego

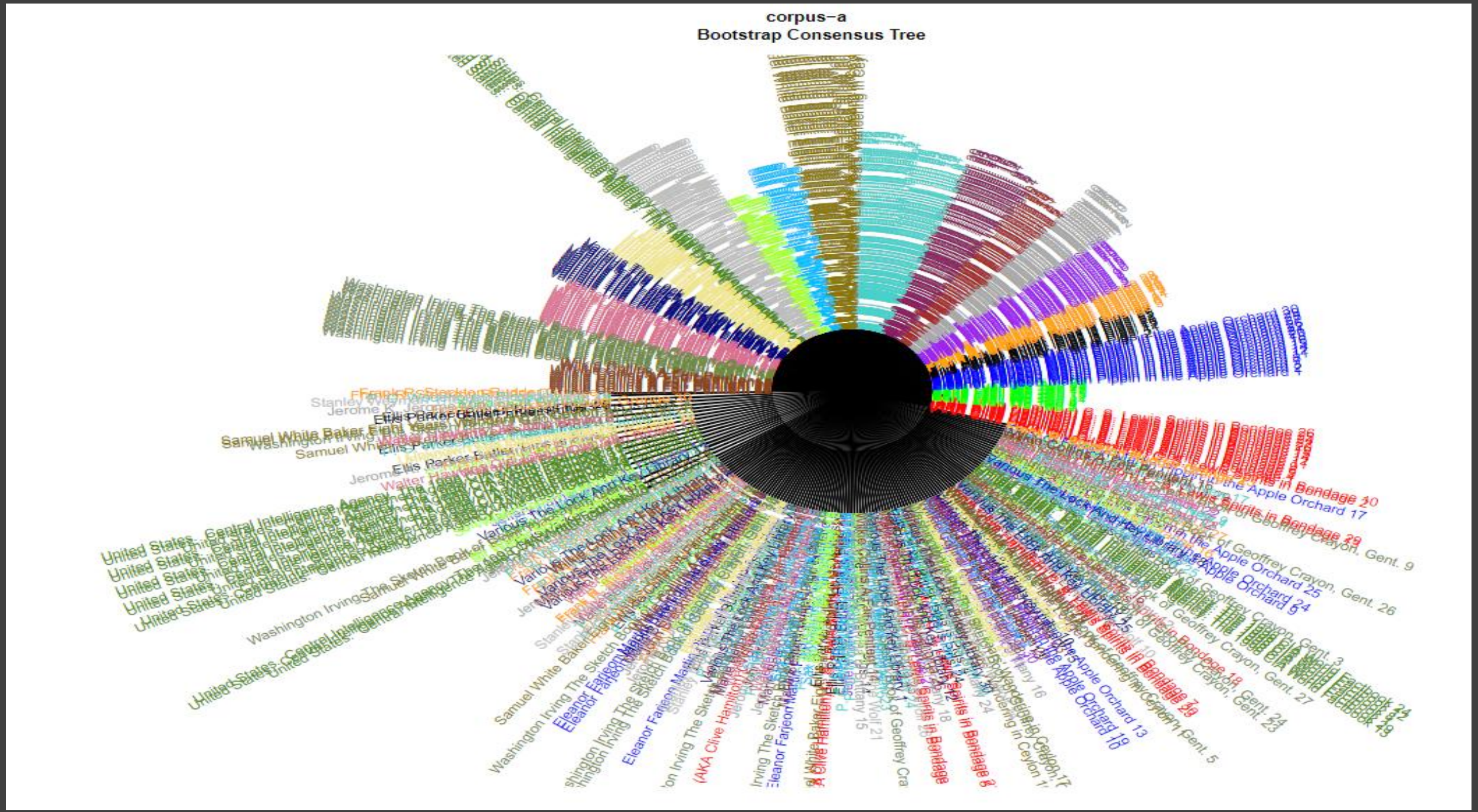
Przykłady działania



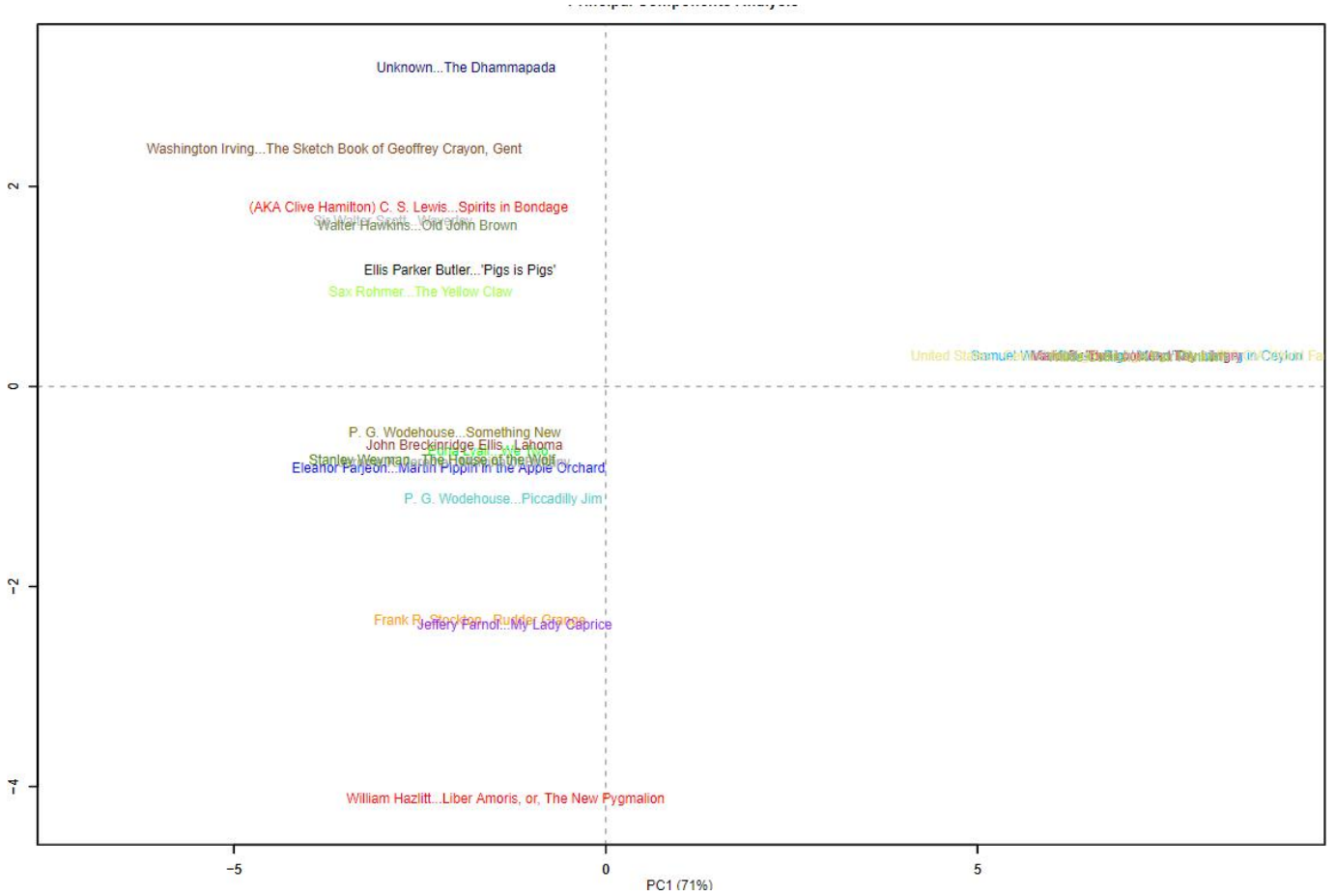
Drzewo zgodności




corpus−a
Bootstrap Consensus Tree
(AKA Clive Hamilton) C. S. Lewis Spirits in Bondage 1
United States. Central Intelligence Agency. The 1998 CIA World Factbook 2
Various The Lock And Key Library
Walter Hawkins Old John Brown 1
Sax Rohmer The Yellow Claw 8
Jeffery Farnol My Lady Caprice 8
Marie Belloc Lowndes The Lodger 10
Samuel White Baker Eight Years' Wandering in Ceylon 9
Geoffrey Crayon, Gent. The Sketch Book of Geoffrey Crayon, Gent. 5
Geoffrey Crayon, Gent. 3
Geoffrey Crayon, Gent. 2
Tipperary Farnol My Lady Caprice 4
Tipperary Farnol My Lady Caprice 7
Tipperary Farnol My Lady Caprice 10
Tipperary Farnol My Lady Caprice 9
Tipperary Farnol My Lady Caprice 6
Tipperary Farnol My Lady Caprice 5
Tipperary Farnol My Lady Caprice 3
Tipperary Farnol My Lady Caprice 2
Tipperary Farnol My Lady Caprice 1
Tipperary Farnol My Lady Caprice 8
Eleanor Farjeon Martin Pippin in the Apple Orchard 10
Eleanor Farjeon Martin Pippin in the Apple Orchard 7
Eleanor Farjeon Martin Pippin in the Apple Orchard 1
Eleanor Farjeon Martin Pippin in the Apple Orchard 9
Eleanor Farjeon Martin Pippin in the Apple Orchard 8
Eleanor Farjeon Martin Pippin in the Apple Orchard 6
Eleanor Farjeon Martin Pippin in the Apple Orchard 5
Eleanor Farjeon Martin Pippin in the Apple Orchard 4
Eleanor Farjeon Martin Pippin in the Apple Orchard 3
Eleanor Farjeon Martin Pippin in the Apple Orchard 2
Marie Belloc Lowndes The Lodger 2
Samuel White Baker Eight Years' Wandering in Ceylon 4
Samuel White Baker Eight Years' Wandering in Ceylon 8
Samuel White Baker Eight Years' Wandering in Ceylon 10
Samuel White Baker Eight Years' Wandering in Ceylon 5


corpus−a
Bootstrap Consensus Tree

PCA


Unknown...The Dhammapada
Washington Irving...The Sketch Book of Geoffrey Crayon, Gent
(AKA Clive Hamilton) C. S. Lewis...Spirits in Bondage
Walter Hawkins...Old John Brown
Ellis Parker Butler...'Pigs is Pigs'
Sax Rohmer...The Yellow Claw
P. G. Wodehouse...Something New
John Breckinridge Ellis...Lahoma
Eleanor Farjeon...Martin Pippin in the Apple Orchard
P. G. Wodehouse...Piccadilly Jim
Jeffery Farnol...My Lady Caprice
William Hazlitt...Liber Amoris, or, The New Pygmalion
2
0
−2
−4
−5
0
5
PC1 (71%)

Stylo również tworzy korzystne pliki ze statystykami, które można wykorzystać dla własnych celów, np…