20.12.2022, Kraków

Analiza Stylu z pomocą Pakietu STYLO w Języku R
Projekt na przedmiot Przetwarzanie Języka Naturalnego

Piotr Gorczowski
Andriei Gensh
Piotr Jurek

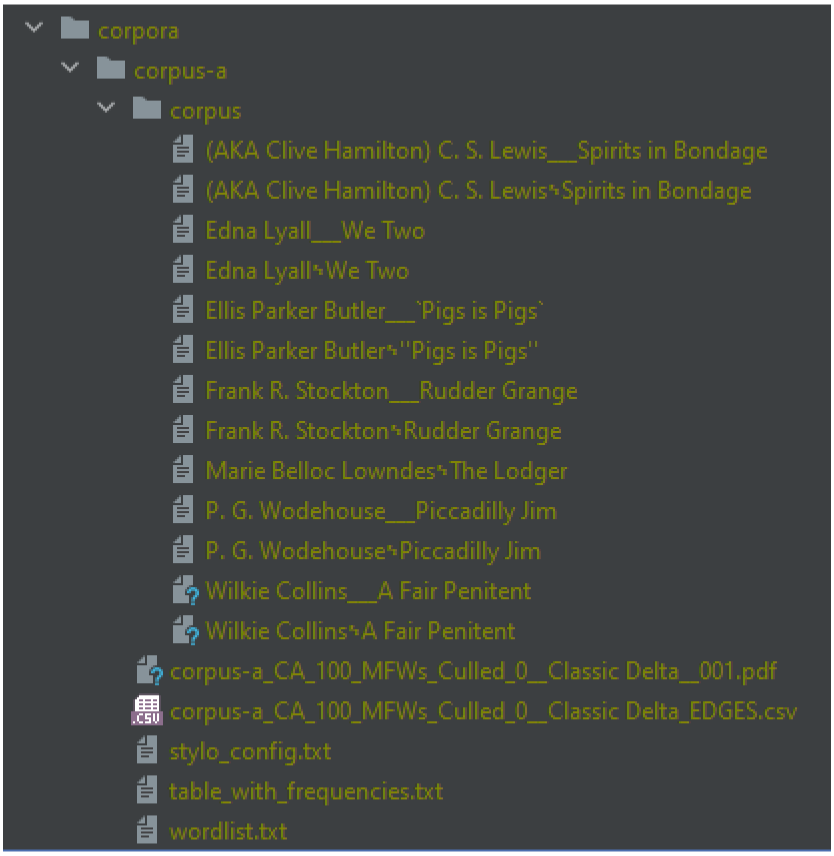
Założenia początkowe projektu:

Celem projektu jest opracowanie i przetestowanie narzędzi do analizy stylów przy użyciu pakietu stylo, takie jak web scraper i prosty program konsolowy w Pythonie do uruchomienia pakietu stylo w odpowiednim położeniu

Przebieg oraz wykonanie projektu:

w celu automatyzacji pracy został wykonany web scraper korzystający z darmowych źródeł https://www.gutenberg.org/ pobierający książki i wyłuskujący z początku tekstu korzystając z podobnego formatowania autora i tytuł książki

program ten został umieszczony w osobnym katalogu book-scraper razem z klasą Book, zapisuje on pobrane pliki w formacie corpora/corpus-a/corpus/author___title.txt gdzie corpora to katalog zawierający korpusy corpus-a to pierwszy rozpatrywany korpus w razie potrzeby można rozpatrywać kolejne np corpus-b ważne aby w jednym katalogu znajdowały się pliki z jedną wersją językową, natomiast pakiet stylo wymaga aby był uruchomiony w katalogu nadrzędnym nad katalogiem “stylo” w którym muszą znajdować się co najmniej dwa źródła.



taka struktura katalogów pozwala również utrzymać porządek w plikach ponieważ wszystkie pliki wynikowe pakiet stylo automatycznie zapisuje do miejsca wykonania czyli w tym przypadku do folderu corpus-a

```python
import requests
import re
import os
from Book import Book

downloaded = 0

FIRST_BOOK_ID = 2000
BOOK_COUNT = 50

for i in range(FIRST_BOOK_ID, FIRST_BOOK_ID+BOOK_COUNT+1):
    url = "https://www.gutenberg.org/cache/epub/{}/pg{}.txt".format(i, i)
    downloadedData = r = requests.get(url)
    decodedData = downloadedData.text
    Author = re.search("(?<=Author: ).*", decodedData)
    Title = re.search("(?<=Title: ).*", decodedData)
    Content = re.search("\*\*\* START OF THIS PROJECT GUTENBERG EBOOK .* \*\*\*.*End of the Project Gutenberg", decodedData, flags=re.DOTALL)

    if (Author != None) and (Title != None) and (Content != None):
        Author = Author.group(0).strip()
        Title = Title.group(0).strip()
        Content = Content.group(0)
        Content = Content[Content.find('\n'):Content.rfind('\n')].strip()
        currentBook = Book(Author, Title, Content)
        currentBook.saveToFile('./corpora/corpus-a/corpus')
        downloaded += 1

print(downloaded)
```

Uruchomienie:

do uruchomienia pakietu stylo wykorzystaliśmy prosty skrypt napisany w języku python przy wykorzystaniu biblioteki RPy2 pozwalającej uruchamiać polecenia R bezpośrednio w kodzia języka python

za pomocą skryptu możemy wybrać i ustawić prawidłową lokalizację roboczą dla działania pakietu stylo domyślnie jest to `r-stylometry/book-scrapper/corpora/corpus-a`

następnie uruchamiamy pakiet stylo i ustawiamy interfejs graficzny na widoczny `gui=True` resztę ustawień można dokonać bezpośrednio w aplikacji okienkowej uruchomionej przez pakiet stylo

Drugim sposobem działania pakietu stylo jest uruchomienie go z poziomu konsoli R jednak wtedy trzeba ręcznie ustawić katalog roboczy przed uruchomieniem pakietu

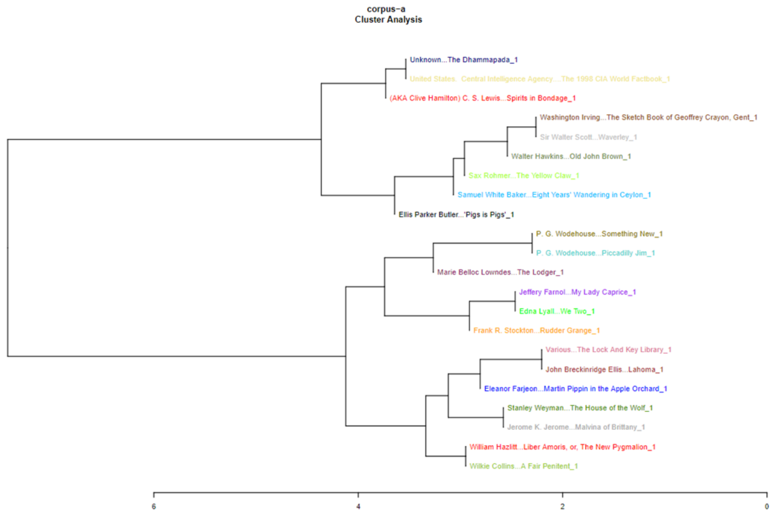
Wykorzystanie Wyników działania programu:

jako wyniki działania programu otrzymujemy pliki:
stylo_config.txt z którego przy kolejnym uruchomieniu wczytywane są ustawienia i zapisywane są tam ustawienia bieżące możemy ten plik również modyfikować przy pomocy skryptu w języku python.

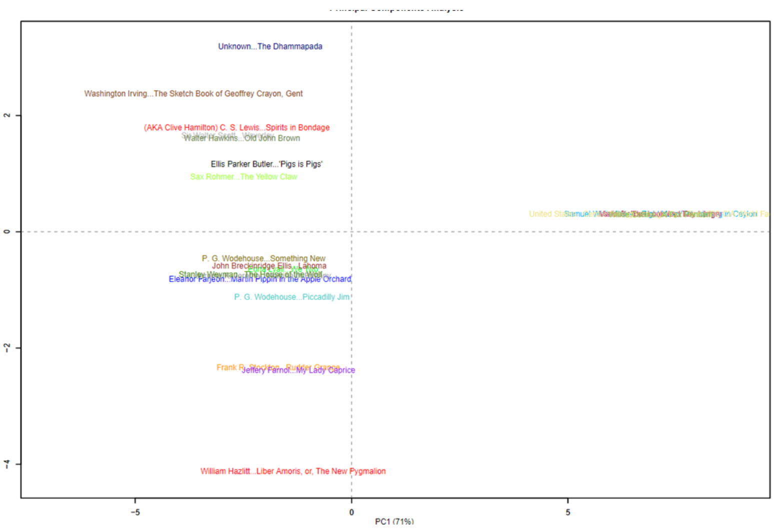
table_with_frequencies.txt plik z listą słów i częstością wystąpień w kolumnach dla każdego analizowanego tekstu.

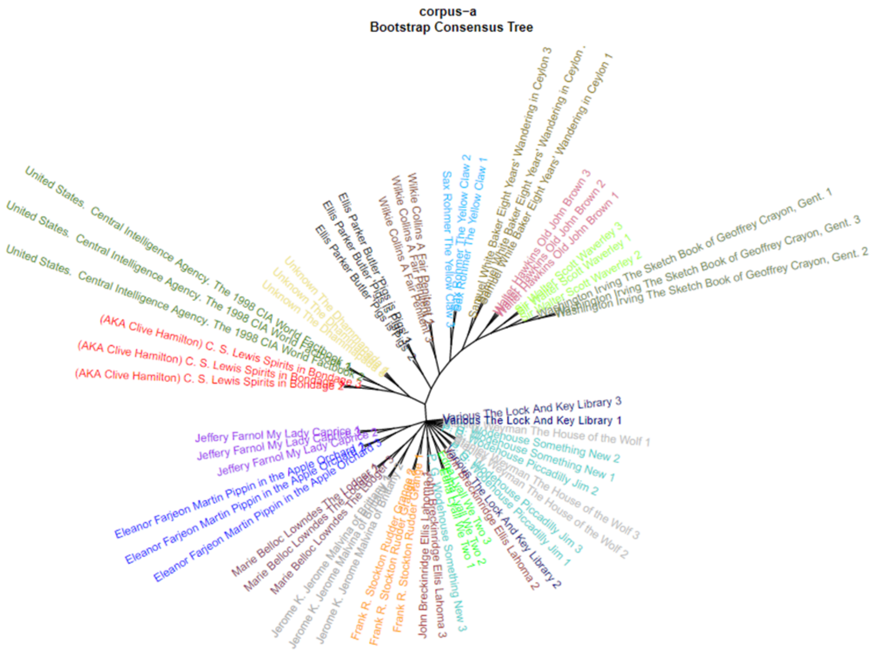
wordlist.txt lista słów

corpus-a_CA_100_MFWs_Culled_0__Classic Delta_EDGES.csv lista porównująca podobieństwo każdego analizowanego pliku z wszystkimi innymi oraz w zależności od ustawień możemy zapisywać różne wynikowe pliki graficzne takie jak graf podobieństw w różnych formatach takich jak PNG PDF itp:



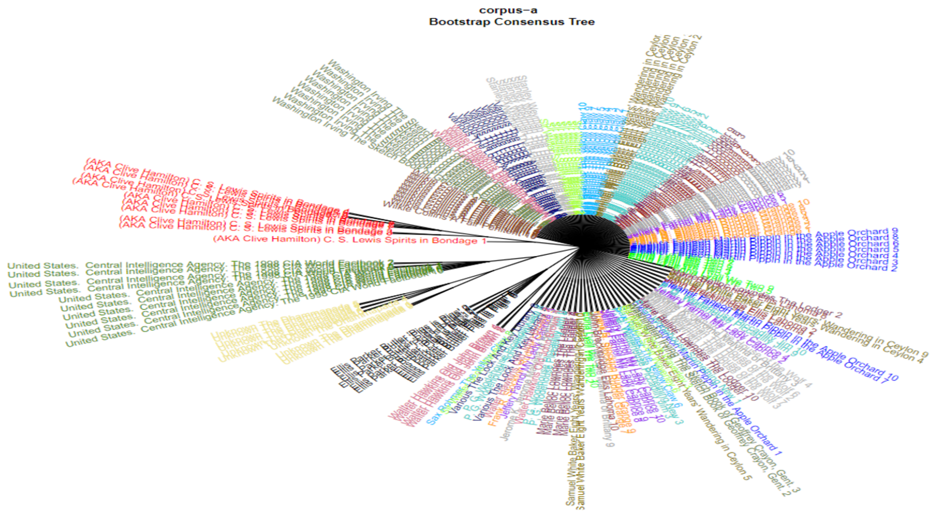
w którym na jednej gałęzi znajdują się dzieła podobne do siebie stylistycznie czy umieszczone utwory w uproszczonej przestrzeni dwuwymiarowej:





jak również możemy generować inne drzewa podobieństw z podziałem tekstu na poszczególne segmenty oznaczone tym samym kolorem jak widać w znaczącej większości przypadków fragmenty tego samego utworu znajdują się w bezpośrednim sąsiedztwie.

w poniższym przykładzie przy dużej ilości tekstów niektóre utwory bardziej mieszają się stylistycznie z innymi



na podstawie plików wynikowych stylo przygotowaliśmy również program wizualizujący wielkością czcionki na wykresie częstość wystąpień danego słowa w badanych materiałach:

Źródła:

http://maciejeder.org/projects/stylo/ - oficjalna strona autora pakietu
http://fizyk.ifpk.pk.edu.pl/~rkycia/events/eRka.html materiały z wykładu na PK
https://rpy2.github.io/ - strona biblioteki do wykorzystania R w pythonie
https://www.gutenberg.org/