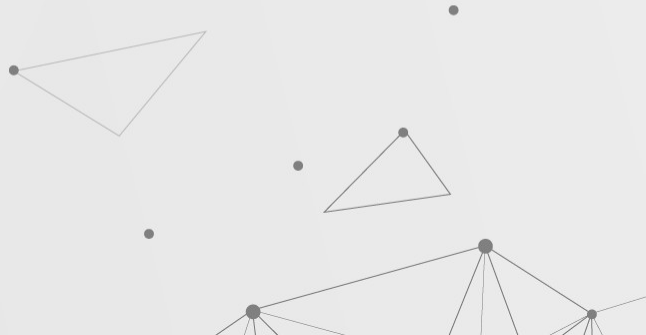
# Narzędzia do stylometrii (analiza stylu artykułów – analiza Autorstwa)

# Biblioteka Stylo

Rozbudowana biblioteka dostępna w języku R stosowana do szerokiej analizy stylistycznej tekstu. Dzięki niej jesteśmy w stanie określić styl autora, a dzięki temu możemy także analizować testy pod kątem autorstwa tekstów.



# Imposters

Funkcja wykorzystująca klasyfikator nadzorowany przez uczenie maszynowe dostosowany do oceny autorstwa. Klasyfikator bazuje na najczęściej występujących słowach.

# Imposters - galbraith

Galbraith - baza danych 3000 najczęściej używanych słów z 26 książek, 5 autorów: Galbraith, Rowling, Cobenm, Tolkien, Lewis.

```
> print(galbraith)
                      the       and       to        of        a         was
coben_breaker       3.5921994 1.1751078 2.1627238 1.3757359 2.5185999 1.5023227
coben_dropshot      3.5878361 1.1785435 2.1221613 1.2685983 2.3753589 1.5674759
coben_fadeaway      3.9313925 1.4454976 2.2004062 1.2130445 2.3064771 1.3225006
coben_falsemove     3.625411  1.6133386 2.1335329 1.2366884 2.4009913 1.3753251
coben_goneforgood   3.8340306 1.816723  2.152941  1.1758076 1.961908  1.7326685
coben_nosecondchance 4.0982934 1.5889666 2.2712554 1.2058633 1.9921373 1.7577145
coben_tellnoone     4.1015556 1.7901359 2.0306373 1.2463421 2.1763598 1.4181288
galbraith_cuckoos   4.5230275 2.267404  2.4940058 2.1793966 2.1412831 1.6555098
lewis_battle        5.0507132 3.4045379 2.1384253 2.1384253 1.9598416 1.5110928
```

# Imposters - galbraith

Sprawdzamy 24 książkę ze zbioru danych, którą jest Drużyna Pierścienia, pod względem autorstwa tekstów. Jak widać najwyższy wynik uzyskał tolkien czyli autor książki, ale styl jest także nieznacznie podobny do lewis'a.

```
> imposters(reference.set=galbraith[-c(24),], test=galbraith[24,])
No candidate set specified; testing the following classes (one at a time):
  coben   galbraith   lewis   rowling   tolkien


Testing a given candidate against imposters...

coben    0
galbraith        0
lewis    0.16
rowling          0
tolkien          1
    coben galbraith     lewis   rowling   tolkien
     0.00      0.00      0.16      0.00      1.00
```

# Imposters - galbraith

Funkcja pozwala konfigurować, w jaki sposób oblicza dystansu, czyli określenie podobieństwa między tekstami. Podstawowym algorytmem jest classic delta, ale można wybrać inny z listy 8 algorytmów np. cosine delta, minmax, argamon. W naszym przypadku nie ma większych różnic, między algorytmami, ale w przypadku minmax czas obliczeń wydłuża się z kilku sekund do kilku minut.

```
    coben galbraith     lewis   rowling   tolkien
     0.00      0.00      0.24      0.00      1.00
> # Cosine Delta (aka Wurzburg Distance)
> imposters(reference.set=galbraith[-c(24),], test=galbraith[24,], distance = "wurzburg")
```

```
    coben galbraith     lewis   rowling   tolkien
      0.0       0.0       0.2       0.0       1.0
> # Ruzicka Distance (aka Minmax Distance)
```

# Imposters - własny korpus

Aby dokonać analizy własnych utworów, należy najpierw wczytać tekst oraz przekształcić go w tokeny. Następnie pobrać najczęściej występujące tokeny. Ostatnim krokiem jest połączenie wyników w jeden macierz.

```
tokenized.texts = load.corpus.and.parse(files = "all")
features = make.frequency.list(tokenized.texts, head = 2000)
data = make.table.of.frequencies(tokenized.texts, features, relative = TRUE)
```

```
> print(data)
                   the       and       to        a         of        he        his
clash-of-kings   5.7233762 2.9413807 2.202715  2.288143  2.1695097 1.7016177 1.4595213
game-of-thrones  5.7294653 2.9664689 2.1966568 2.2235359 2.0301196 1.7630208 1.5286281
hamlet           3.3483857 2.9920437 2.3346543 1.6987682 2.0858293 0.8202009 0.9430774
king-lear        2.9222967 2.6456715 2.1278859 1.4824272 1.6845764 0.8511544 0.7447601
macbeth          3.5940011 3.0637386 2.126406  1.3604713 1.8211034 0.7445099 0.803428
romeo-and-juliet 2.5905366 2.7189306 2.0354216 1.7106605 1.4538726 0.6344171 0.5400098
```

# Imposters - własny korpus

Wczytane zostały 6 książek 2 autorów: George R. R. Martin (Clash Of Kings, A Game of Thrones), William Shakespeare (Macbeth, Hamlet, Romeo And Juliet, King Lear).

Dokonujemy analizę autorstwa książki Clash Of Kings w porównaniu do pozostałych książek. Wynik jest taki jaki oczekiwaliśmy czyli podobieństwo 1.0 do drugiej książki tego samego autora.

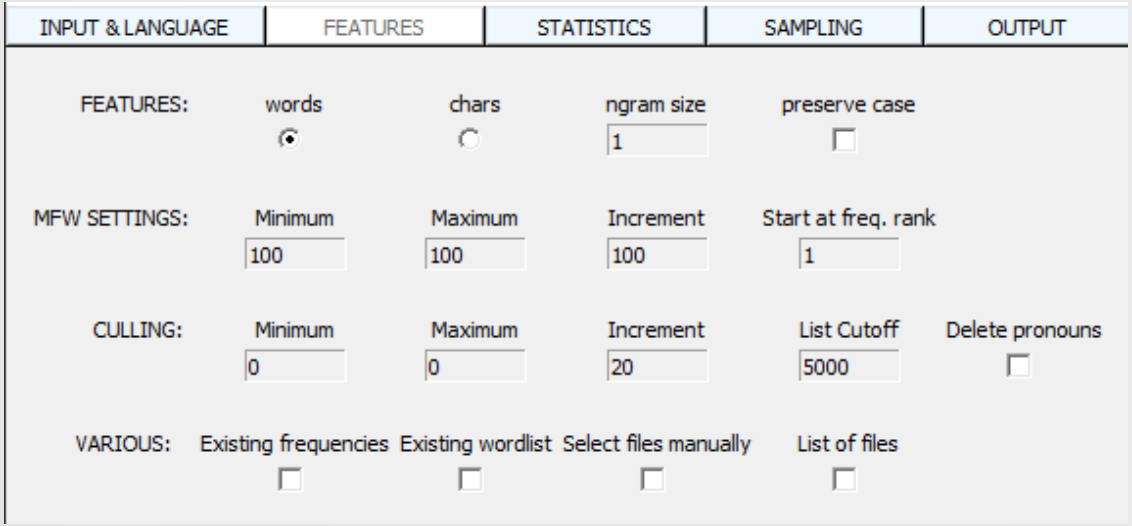
| game-of-thrones | hamlet | king-lear | macbeth | romeo-and-juliet |
|---|---|---|---|---|
| 1.00 | 0.23 | 0.19 | 0.21 | 0.00 |

# Stylo

Wywołując funkcję stylo zostanie otwarta aplikacja w trybie graficznym, która pozwala w prosty sposób konfigurować ustawienia na postawie, których chcemy dokonać analizy stylu artysty. W aplikacji jest wiele opcji do konfiguracji, z czego najważniejsze są zakładki FEATURES, STATISTICS, SAMPLING.

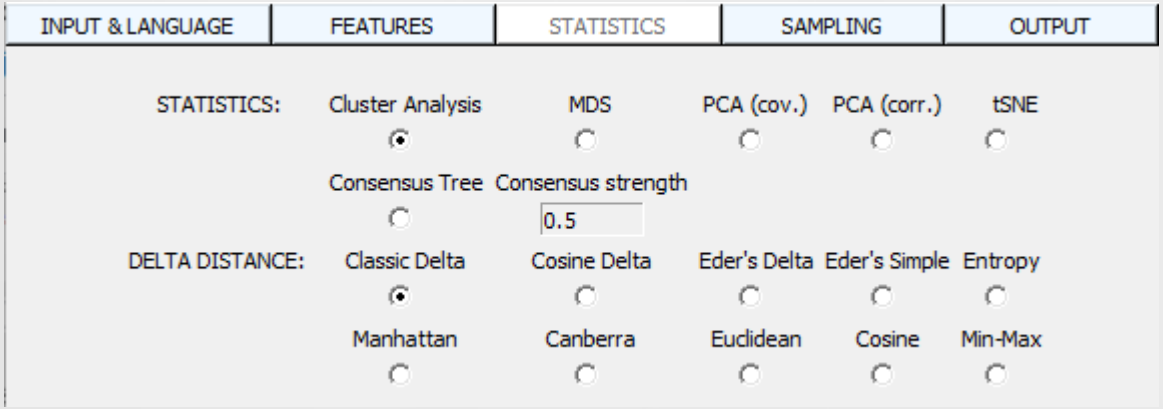
# Stylo - FEATURES

Opcje w zakładce pozwalają konfigurować, na jakiej zasadzie tworzone są cechy, ile ich ma być oraz możliwość na pomijanie pewnych cech, aby zwiększyć skuteczność.



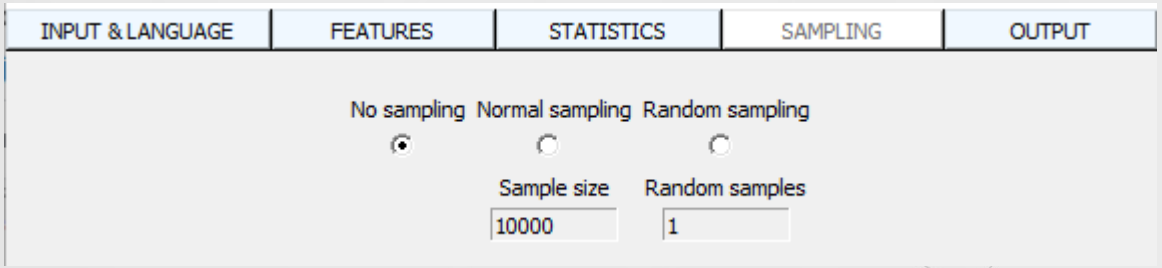
# Stylo - STATISTICS

Opcje w zakładce pozwalają wybór, jaką statystykę chcemy sprawdzić oraz algorytm obliczania podobieństwa między tekstami.



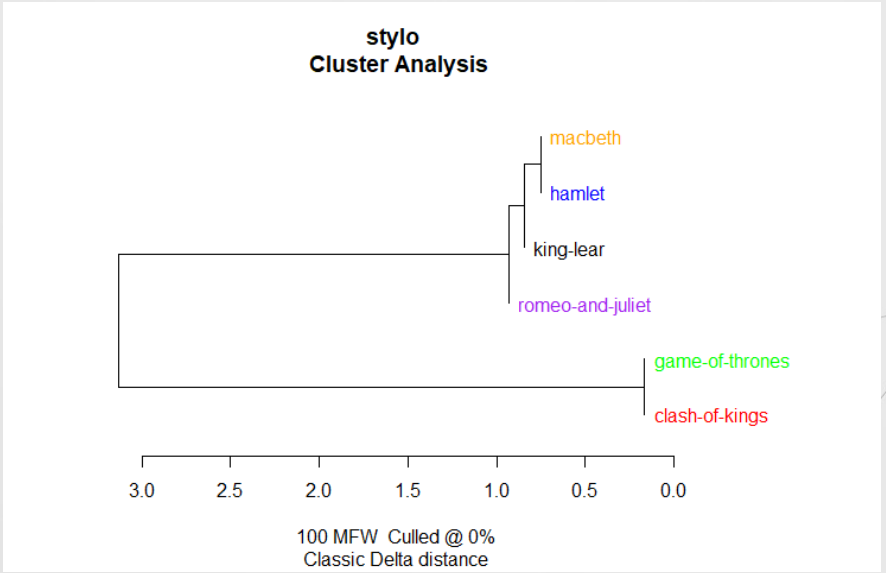
# Stylo - SAMPLING

Zakładka sampling pozwala na konfigurację podziału analizy tekstu na mniejsze fragmenty w przypadku, gdyby tekst był za duży do analizy.



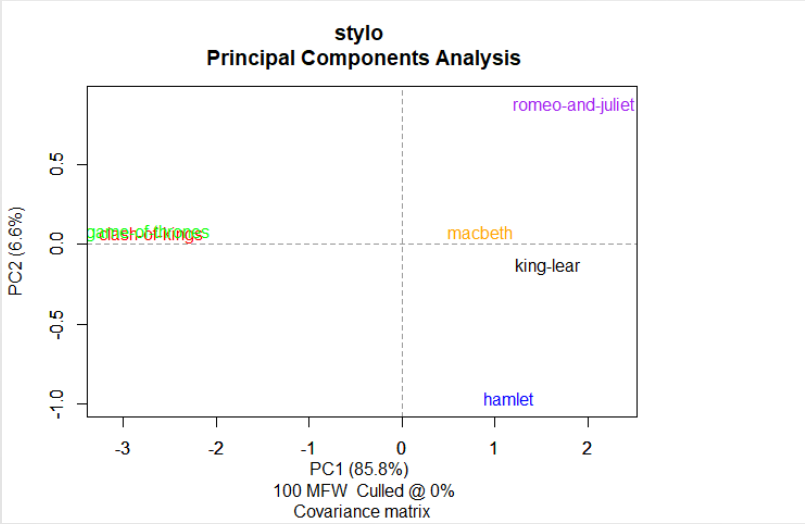
# Stylo - Cluster Analysis

Cluster Analysis - analiza, która ma na celu wyznaczenie grup na podstawie pewnych cech. W naszym przypadku książki tego samego autora zostały połączone w tę samą grupę.

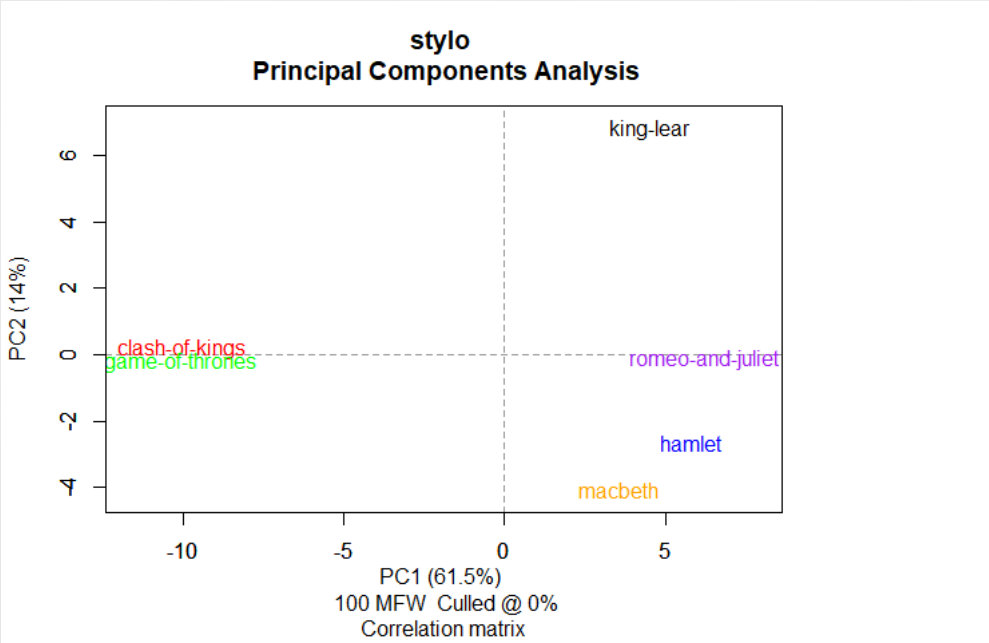


# Stylo - PCA (covariance)

PCA - analiza głównych składowych, która ma na celu wyznaczenie nowej przestrzeni obserwacji, w której najwięcej zmienności wyjaśniają początkowe czynniki. PCA może być oparta o macierz kowariancji lub macierz korelacji.



# Stylo - PCA (correlation)



# Rolling Classify

Funkcja dzieląca tekst na kolejne bloki o równej wielkości i przeprowadzająca nadzorowaną klasyfikację tych bloków względem zbioru uczącego. Funkcja ta może być wykorzystywana między innymi do tego, aby sprawdzać, które części tekstu mają inny styl co może oznaczać, że zostały napisane przez innego autora.

```
rolling.classify(
  write.png.file = TRUE,
  classification.method = "svm",
  mfw = 100,
  training.set.sampling = "normal.sampling",
  slice.size = 5000,
  slice.overlap = 4500
)
```

# Rolling Classify

Do przetestowania zostały wykorzystane 3 książki:

- Good Omens (Neil Gaiman, Terry Pratchett)
- Small Gods (Terry Pratchett)
- The Graveyard Book (Neil Gaiman)

Good Omens zostało napisane przez dwóch autorów i chcemy dowiedzieć się, jak autorzy podzielili się w czasie tworzenia dzieła.

# Rolling Classify