# PROJEKT 2 MILESTONE 2

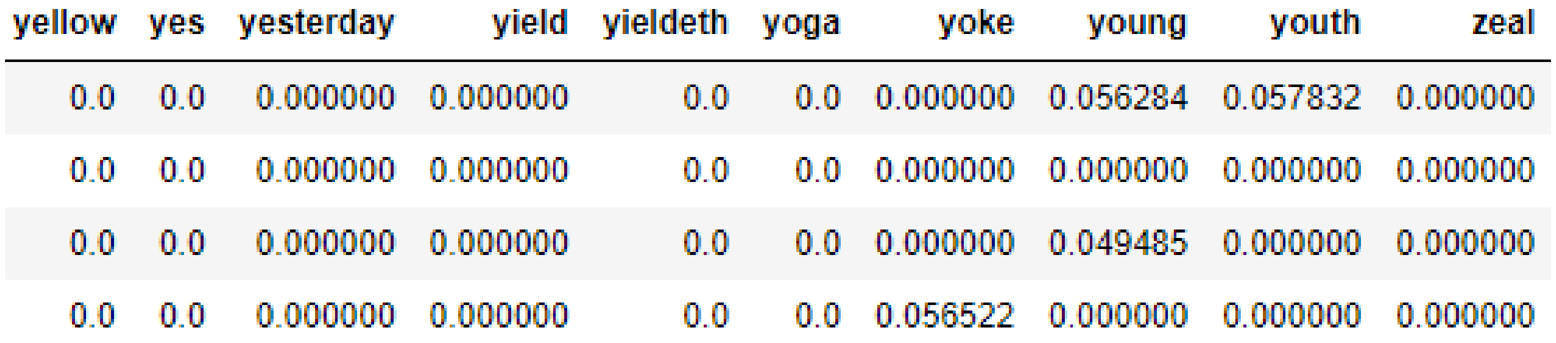
Przemysław Olender, Dominik Pawlak

# PRZYGOTOWANIE DANYCH - TF IDF, PRZESKALOWANIE

- Stworzyliśmy również ramkę z wykorzystaniem narzędzia TF DIF.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad idf(w) = log(\frac{N}{df_t}) \qquad w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

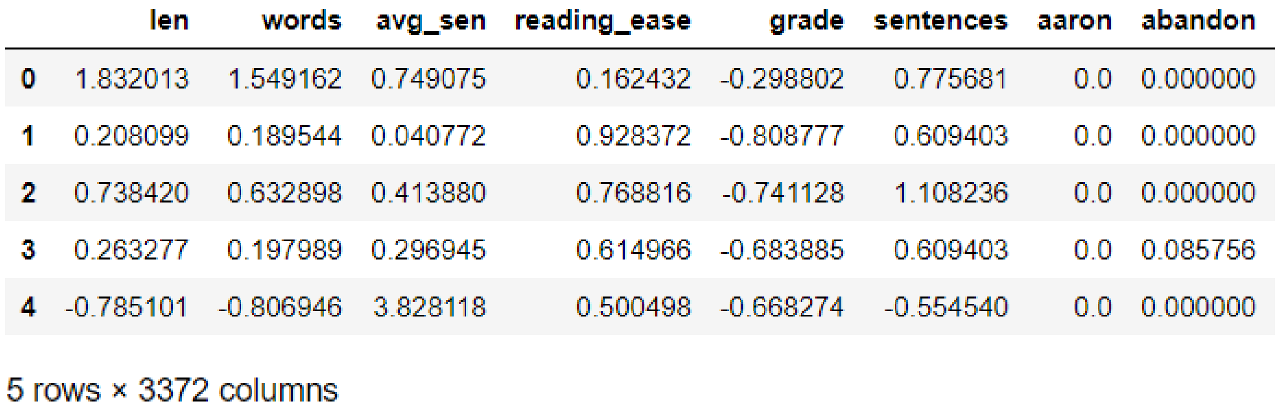
- Otrzymaliśmy następującą ramkę danych



| yellow | yes | yesterday | yield | yieldeth | yoga | yoke | young | youth | zeal |
|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.056284 | 0.057832 | 0.000000 |
| 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.049485 | 0.000000 | 0.000000 |
| 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.056522 | 0.000000 | 0.000000 | 0.000000 |

# PRZYGOTOWANIE DANYCH - SKALOWANIE

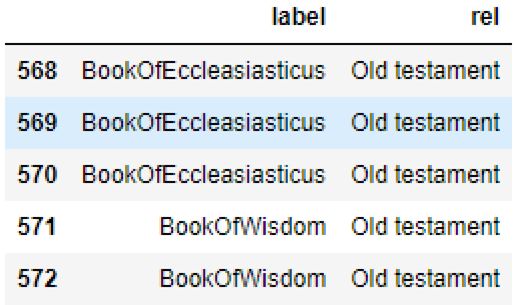
- Za pomocą Standard Scalera przeskalowaliśmy ramkę ze statystykami.



| | len | words | avg_sen | reading_ease | grade | sentences | aaron | abandon |
|---|---|---|---|---|---|---|---|---|
| **0** | 1.832013 | 1.549162 | 0.749075 | 0.162432 | -0.298802 | 0.775681 | 0.0 | 0.000000 |
| **1** | 0.208099 | 0.189544 | 0.040772 | 0.928372 | -0.808777 | 0.609403 | 0.0 | 0.000000 |
| **2** | 0.738420 | 0.632898 | 0.413880 | 0.768816 | -0.741128 | 1.108236 | 0.0 | 0.000000 |
| **3** | 0.263277 | 0.197989 | 0.296945 | 0.614966 | -0.683885 | 0.609403 | 0.0 | 0.085756 |
| **4** | -0.785101 | -0.806946 | 3.828118 | 0.500498 | -0.668274 | -0.554540 | 0.0 | 0.000000 |

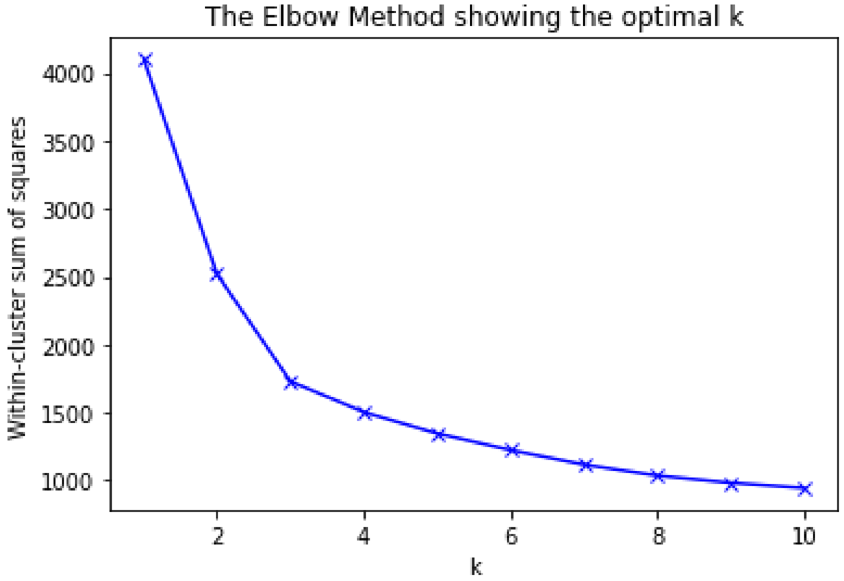
5 rows × 3372 columns

- Stworzyliśmy też ramkę z odpowiedzi, pogrupowaliśmy teskty według religii

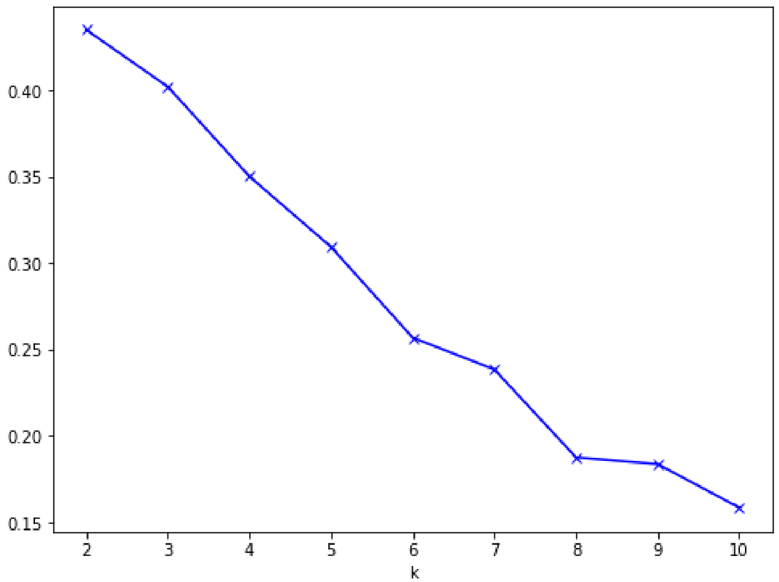


| | label | rel |
|---|---|---|
| **568** | BookOfEccleasiasticus | Old testament |
| **569** | BookOfEccleasiasticus | Old testament |
| **570** | BookOfEccleasiasticus | Old testament |
| **571** | BookOfWisdom | Old testament |
| **572** | BookOfWisdom | Old testament |

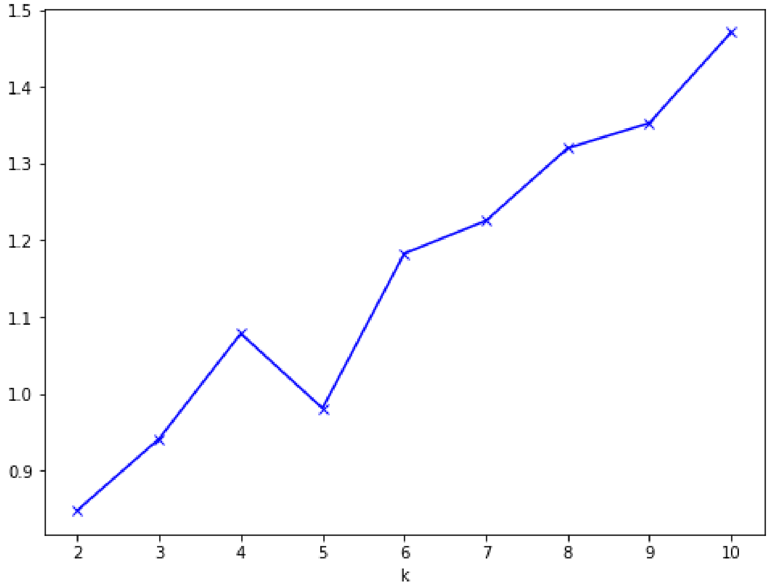
# WYBÓR LICZBY KLASTRÓW
## (NA PEŁNYM ZBIORZE)



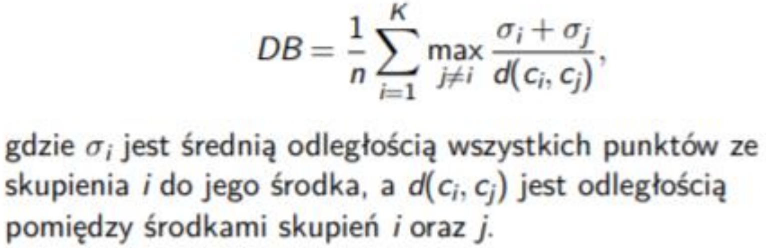
Metryka łokciowa



Metryka Silhouette



Metryka Davies Bouldin



$$DB = \frac{1}{n}\sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)},$$

gdzie $\sigma_i$ jest średnią odległością wszystkich punktów ze skupienia $i$ do jego środka, a $d(c_i, c_j)$ jest odległością pomiędzy środkami skupień $i$ oraz $j$.

# METRYKI

- Silhouette score
- Davies bouldin score
- Rand score

$$RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}}$$

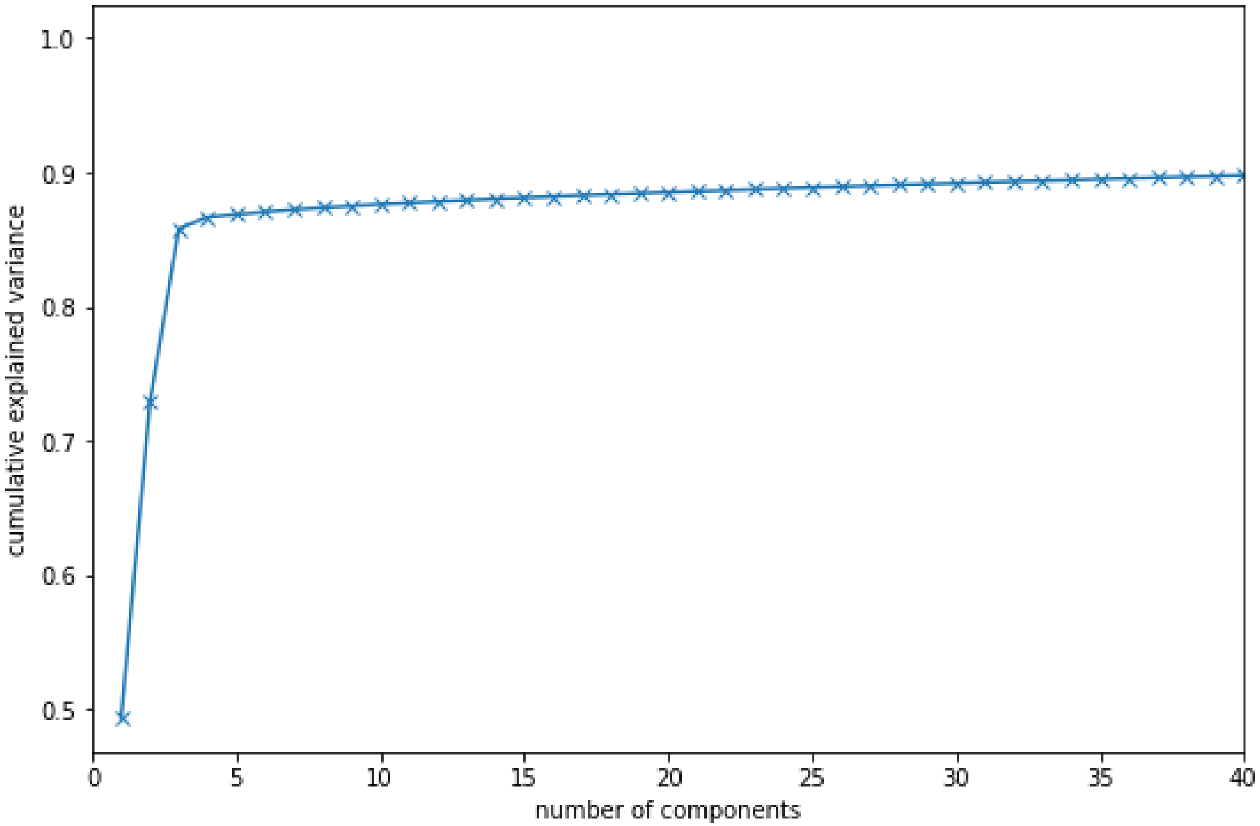
- Adjusted mutual info score

$$ARI = \frac{\text{RI - Expected RI}}{\text{Max(RI) - Expected RI}}$$

- Mutual info score

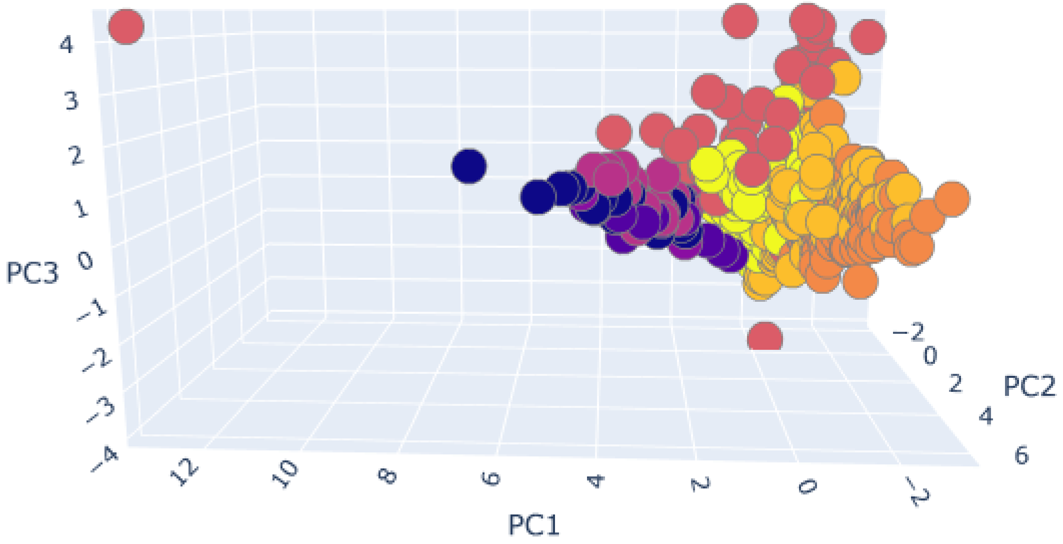
$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

# PCA – SKUMULOWANA WARIANCJA

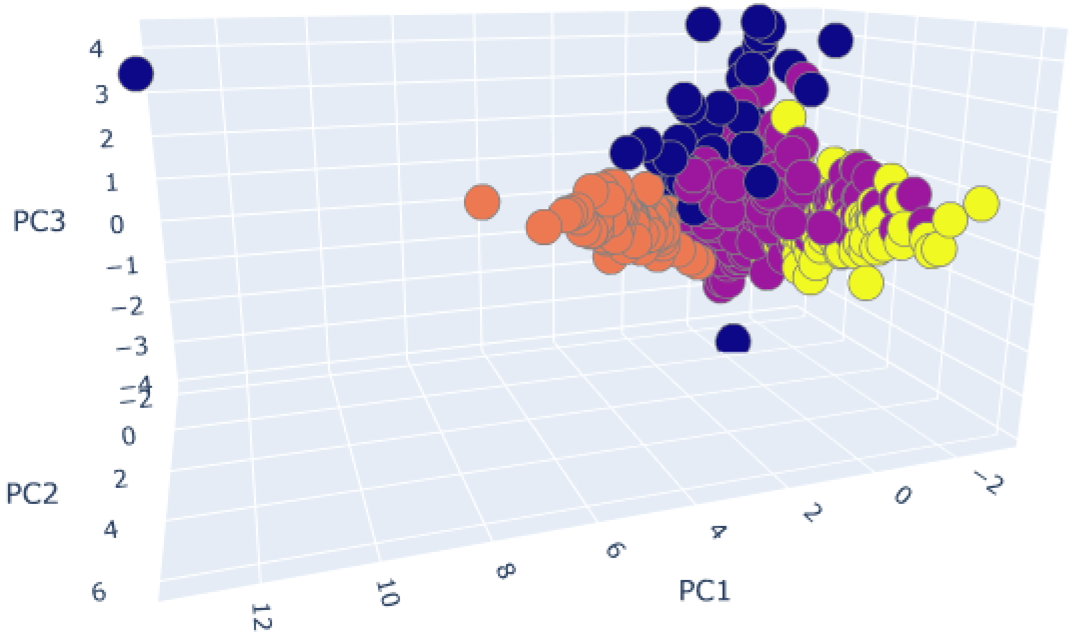


# PCA DLA 3 KOMPONENTÓW

## (ZBIÓR Z ETYKIETAMI)



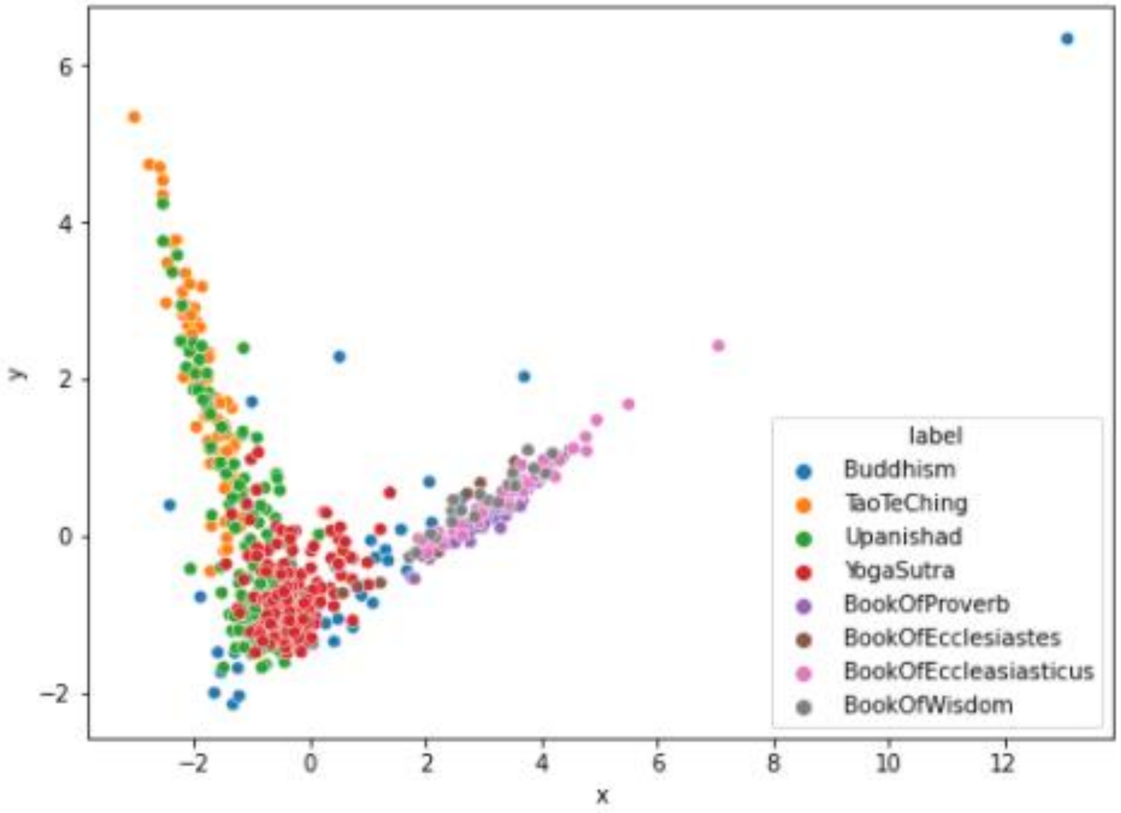
Podział na teksty

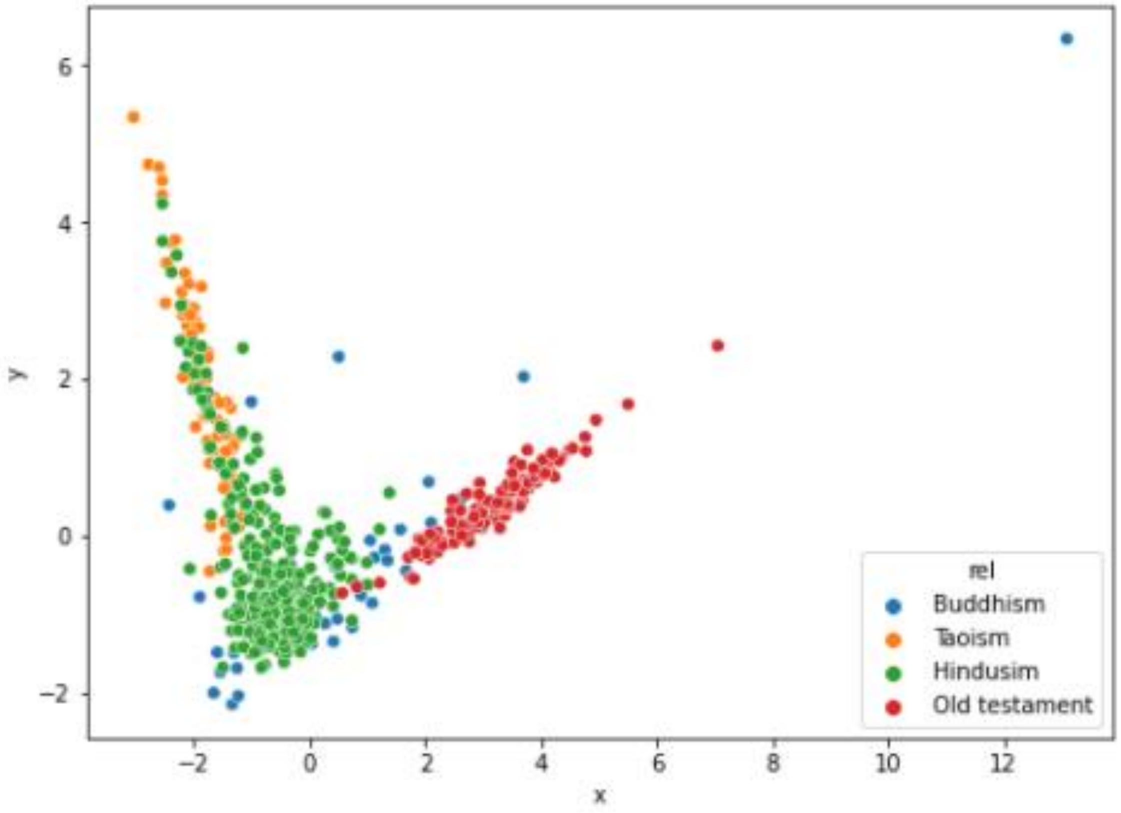


Podział na religie

# PCA DLA 2 KOMPONENTÓW

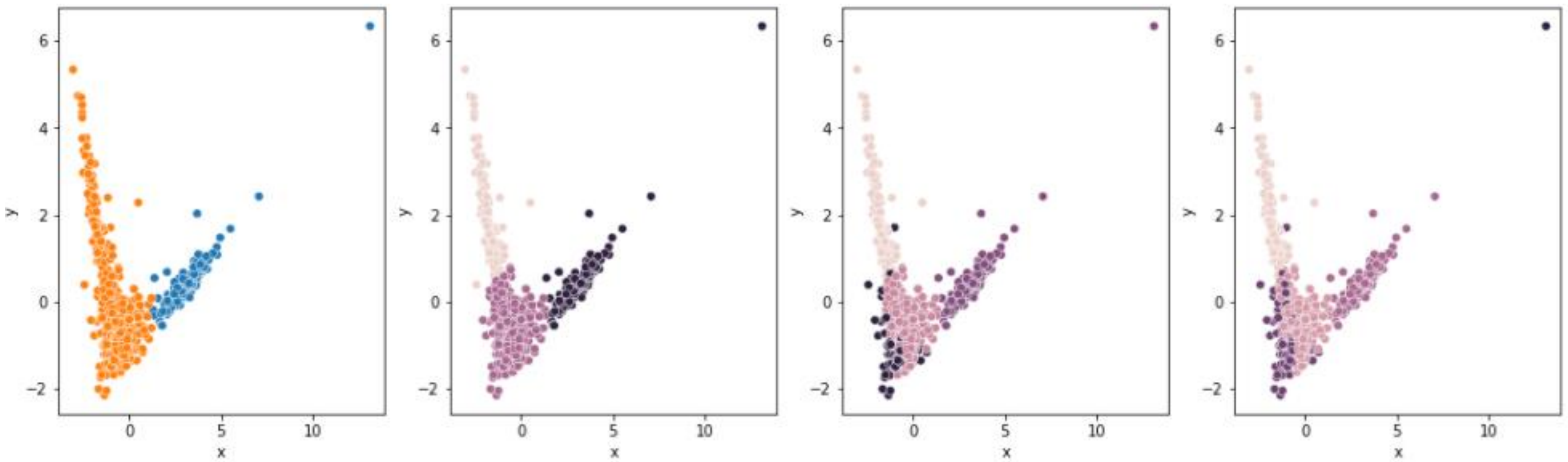
## (ZBIÓR Z ETYKIETAMI)

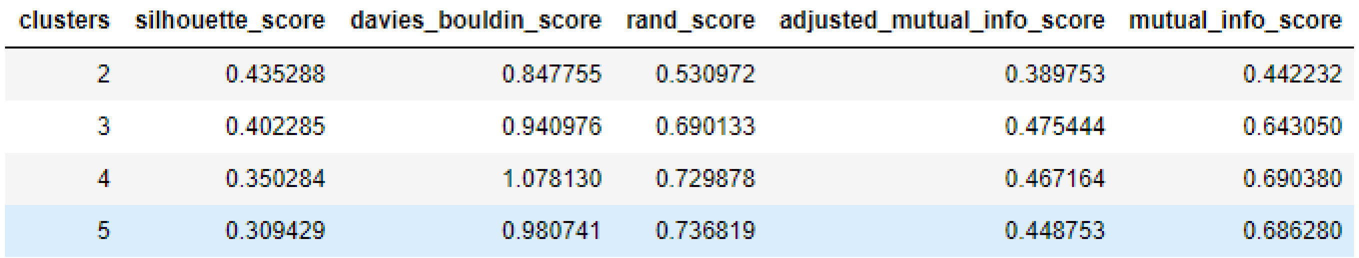




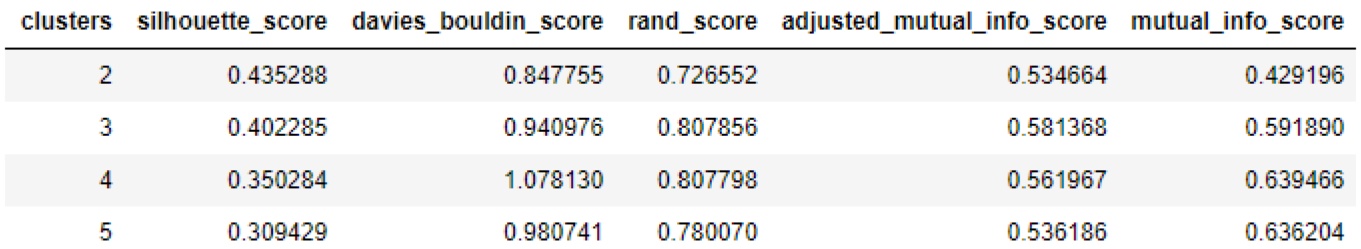
# K-MEANS BEZ REDUKCJI WYMIARÓW

## WIZUALIZACJA PO PCA





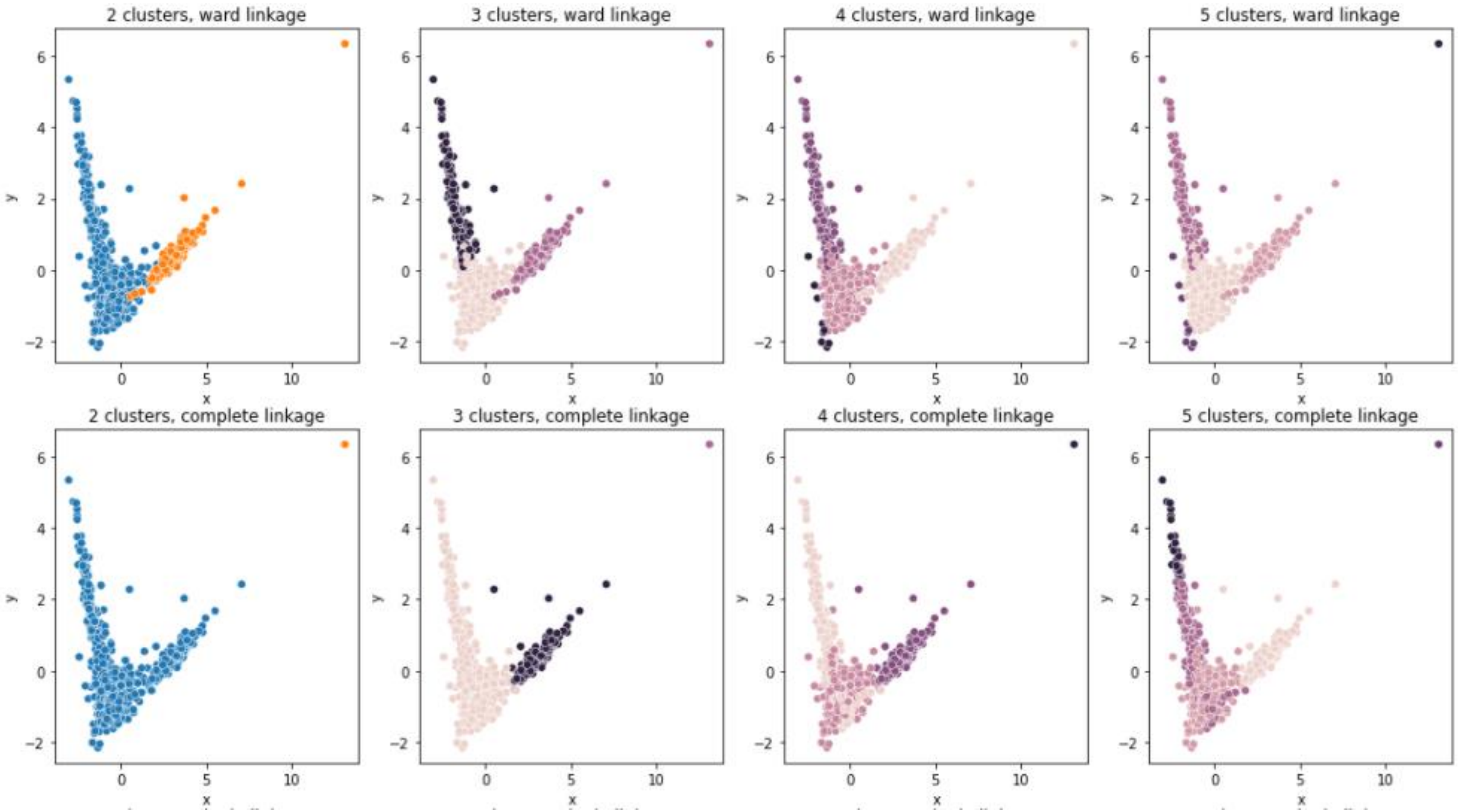
| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.435288 | 0.847755 | 0.530972 | 0.389753 | 0.442232 |
| 3 | 0.402285 | 0.940976 | 0.690133 | 0.475444 | 0.643050 |
| 4 | 0.350284 | 1.078130 | 0.729878 | 0.467164 | 0.690380 |
| 5 | 0.309429 | 0.980741 | 0.736819 | 0.448753 | 0.686280 |



| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.435288 | 0.847755 | 0.726552 | 0.534664 | 0.429196 |
| 3 | 0.402285 | 0.940976 | 0.807856 | 0.581368 | 0.591890 |
| 4 | 0.350284 | 1.078130 | 0.807798 | 0.561967 | 0.639466 |
| 5 | 0.309429 | 0.980741 | 0.780070 | 0.536186 | 0.636204 |

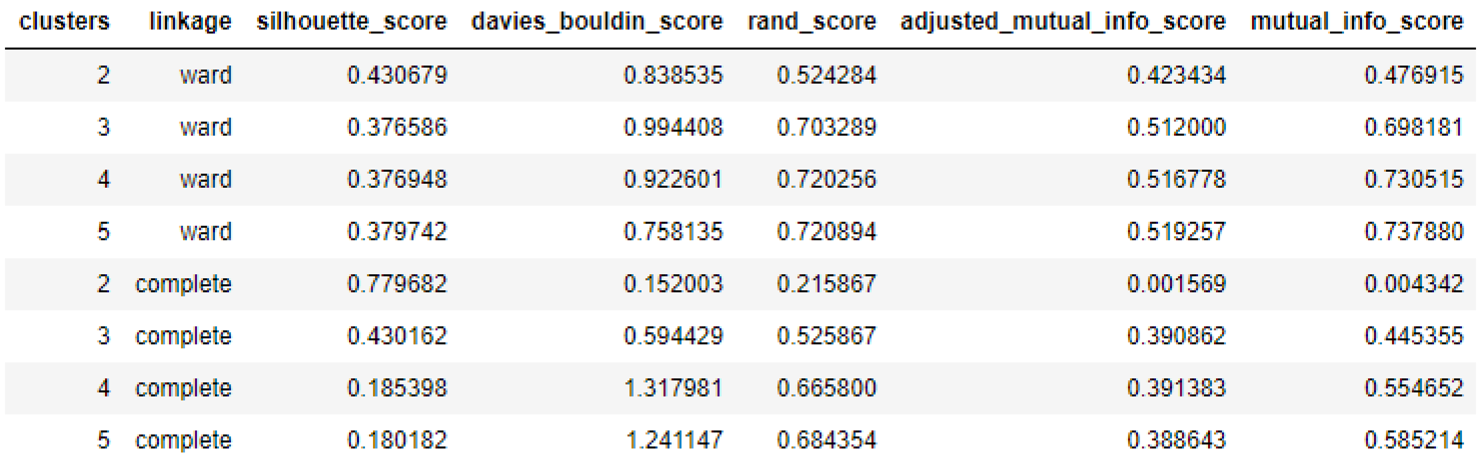
# AGGLOMERATIVE CLUSTERING

## WIZUALIZACJA PO PCA, BEZ REDUKCJI WYMIARÓW

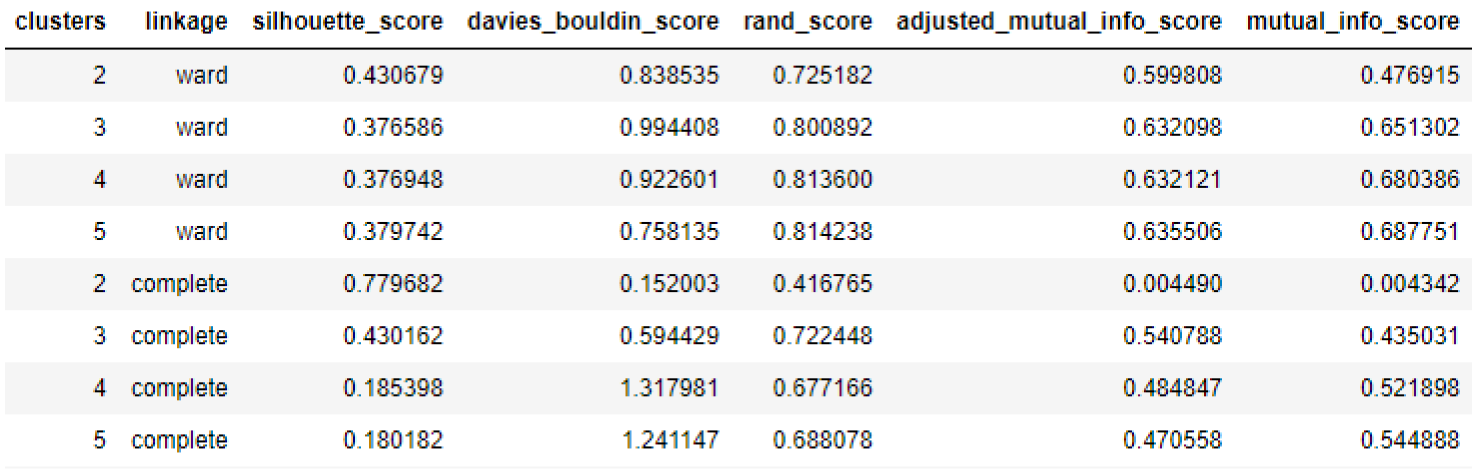


# AGGLOMERATIVE CLUSTERING

## WIZUALIZACJA PO PCA



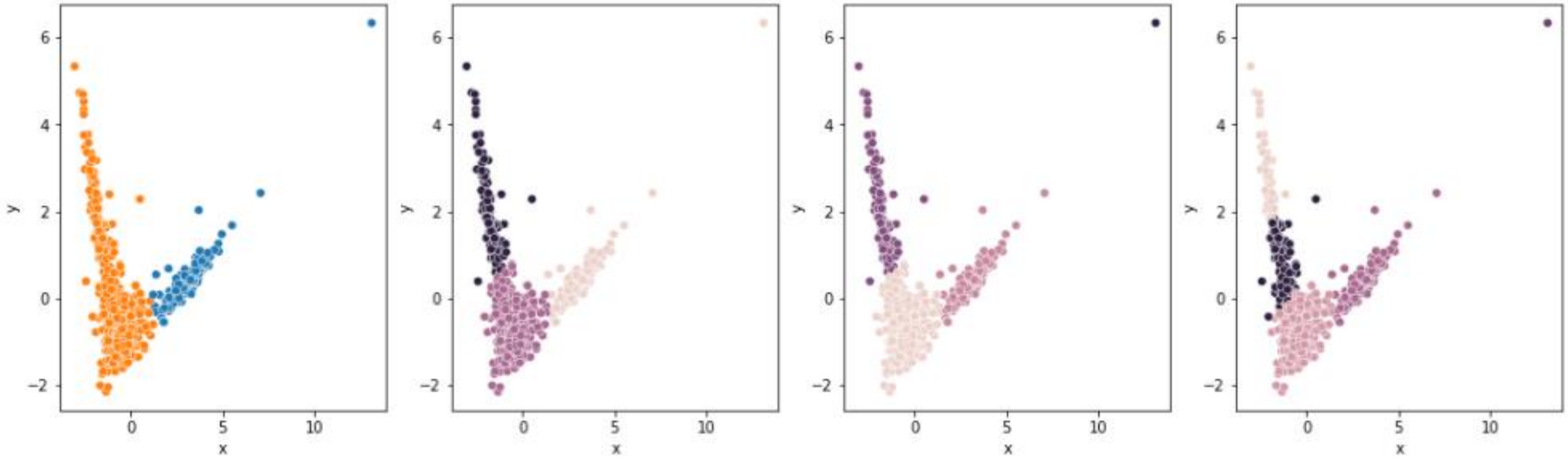
| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.430679 | 0.838535 | 0.524284 | 0.423434 | 0.476915 |
| 3 | ward | 0.376586 | 0.994408 | 0.703289 | 0.512000 | 0.698181 |
| 4 | ward | 0.376948 | 0.922601 | 0.720256 | 0.516778 | 0.730515 |
| 5 | ward | 0.379742 | 0.758135 | 0.720894 | 0.519257 | 0.737880 |
| 2 | complete | 0.779682 | 0.152003 | 0.215867 | 0.001569 | 0.004342 |
| 3 | complete | 0.430162 | 0.594429 | 0.525867 | 0.390862 | 0.445355 |
| 4 | complete | 0.185398 | 1.317981 | 0.665800 | 0.391383 | 0.554652 |
| 5 | complete | 0.180182 | 1.241147 | 0.684354 | 0.388643 | 0.585214 |

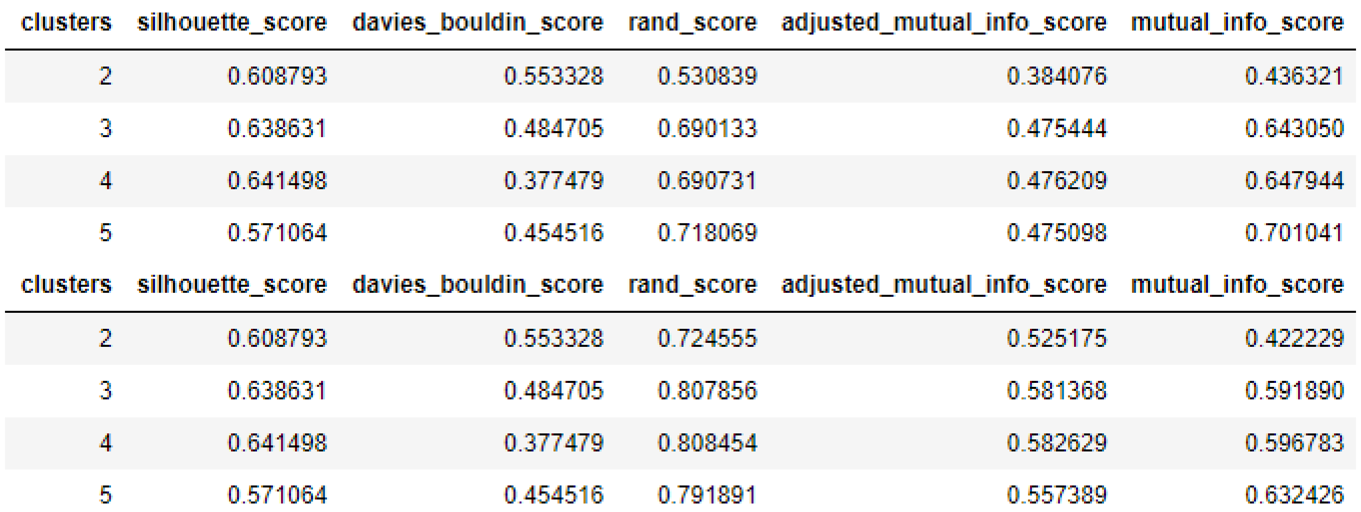


| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.430679 | 0.838535 | 0.725182 | 0.599808 | 0.476915 |
| 3 | ward | 0.376586 | 0.994408 | 0.800892 | 0.632098 | 0.651302 |
| 4 | ward | 0.376948 | 0.922601 | 0.813600 | 0.632121 | 0.680386 |
| 5 | ward | 0.379742 | 0.758135 | 0.814238 | 0.635506 | 0.687751 |
| 2 | complete | 0.779682 | 0.152003 | 0.416765 | 0.004490 | 0.004342 |
| 3 | complete | 0.430162 | 0.594429 | 0.722448 | 0.540788 | 0.435031 |
| 4 | complete | 0.185398 | 1.317981 | 0.677166 | 0.484847 | 0.521898 |
| 5 | complete | 0.180182 | 1.241147 | 0.688078 | 0.470558 | 0.544888 |

# K-MEANS PO PCA

## WIZUALIZACJA PO PCA

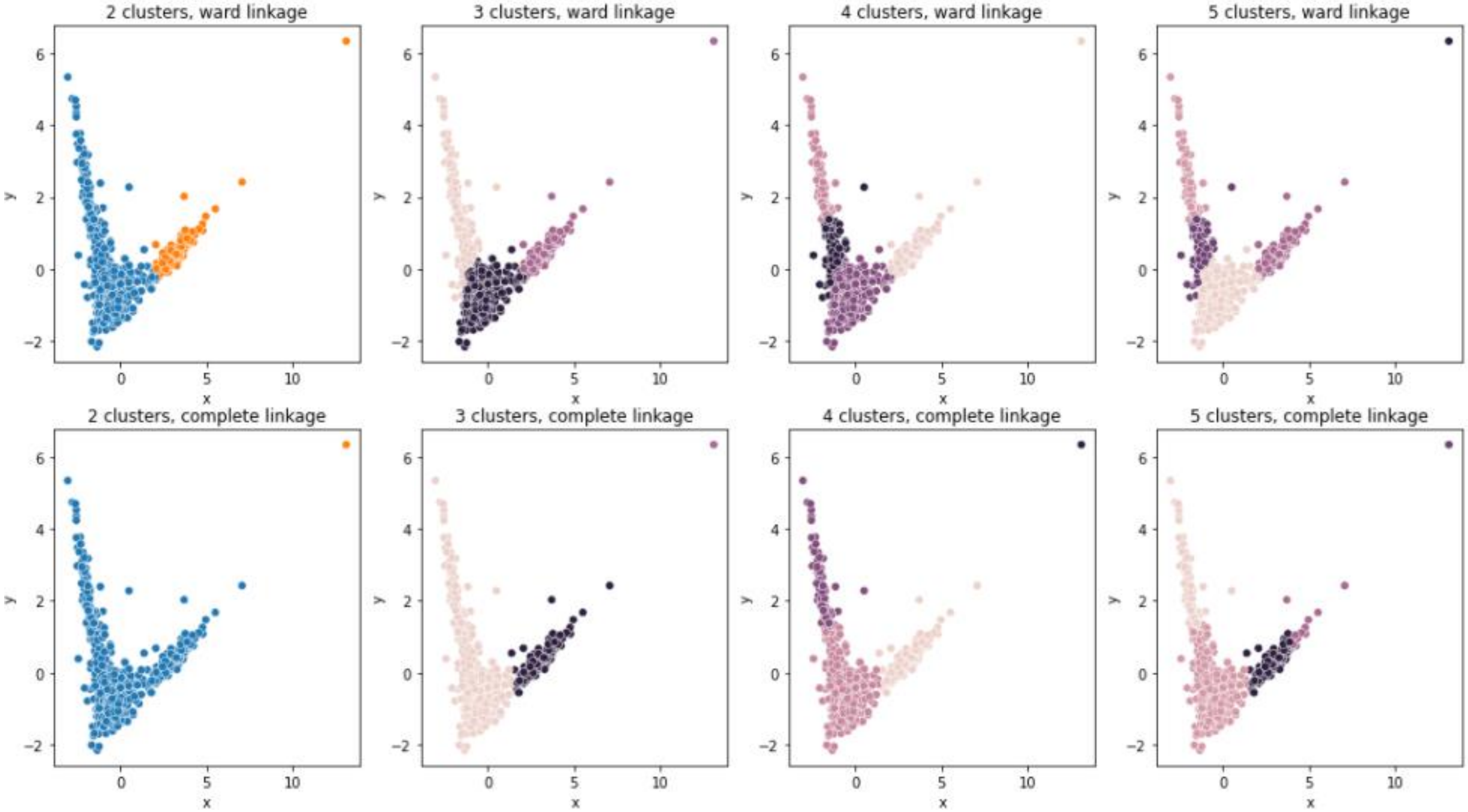




| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.608793 | 0.553328 | 0.530839 | 0.384076 | 0.436321 |
| 3 | 0.638631 | 0.484705 | 0.690133 | 0.475444 | 0.643050 |
| 4 | 0.641498 | 0.377479 | 0.690731 | 0.476209 | 0.647944 |
| 5 | 0.571064 | 0.454516 | 0.718069 | 0.475098 | 0.701041 |

| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.608793 | 0.553328 | 0.724555 | 0.525175 | 0.422229 |
| 3 | 0.638631 | 0.484705 | 0.807856 | 0.581368 | 0.591890 |
| 4 | 0.641498 | 0.377479 | 0.808454 | 0.582629 | 0.596783 |
| 5 | 0.571064 | 0.454516 | 0.791891 | 0.557389 | 0.632426 |

# AGGLOMERATIVE CLUSTERING PO PCA



2 clusters, ward linkage

3 clusters, ward linkage

4 clusters, ward linkage

5 clusters, ward linkage

2 clusters, complete linkage

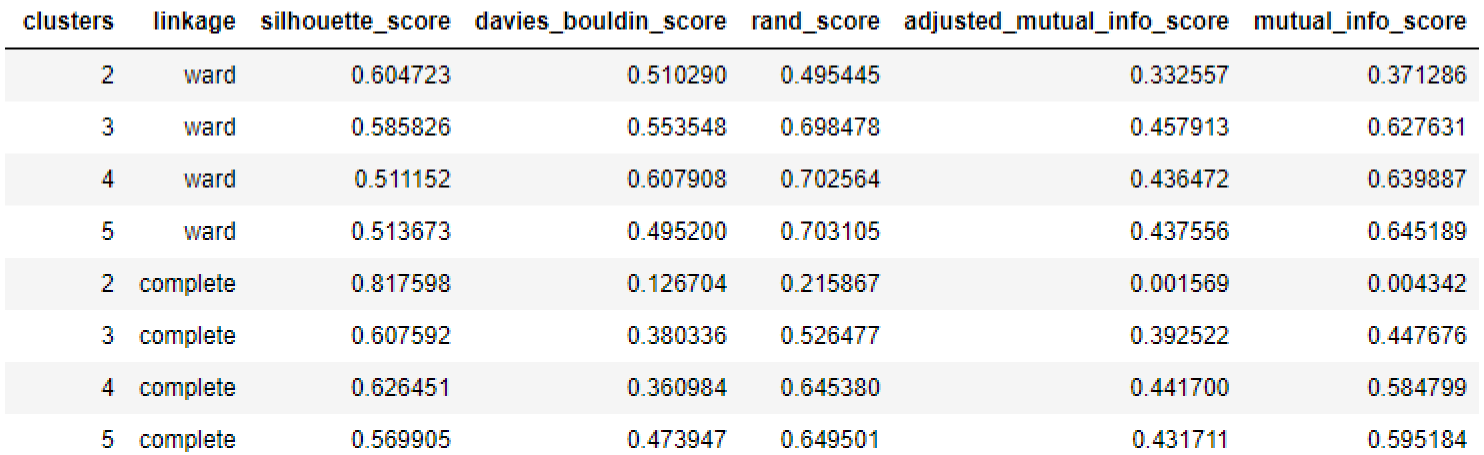
3 clusters, complete linkage

4 clusters, complete linkage

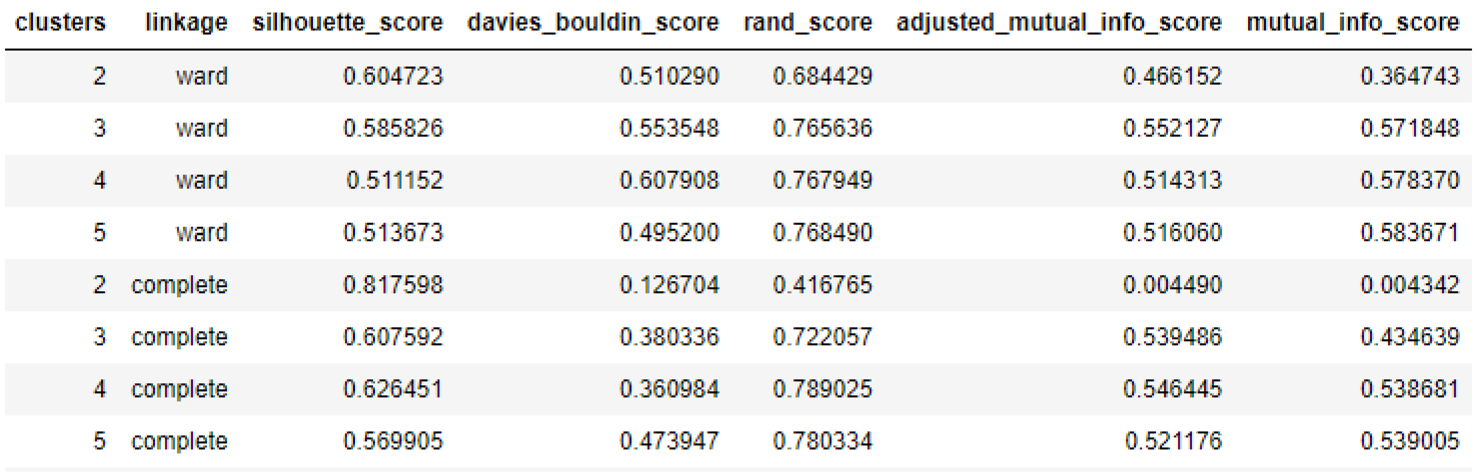
5 clusters, complete linkage

# AGGLOMERATIVE CLUSTERING

## WIZUALIZACJA PO PCA

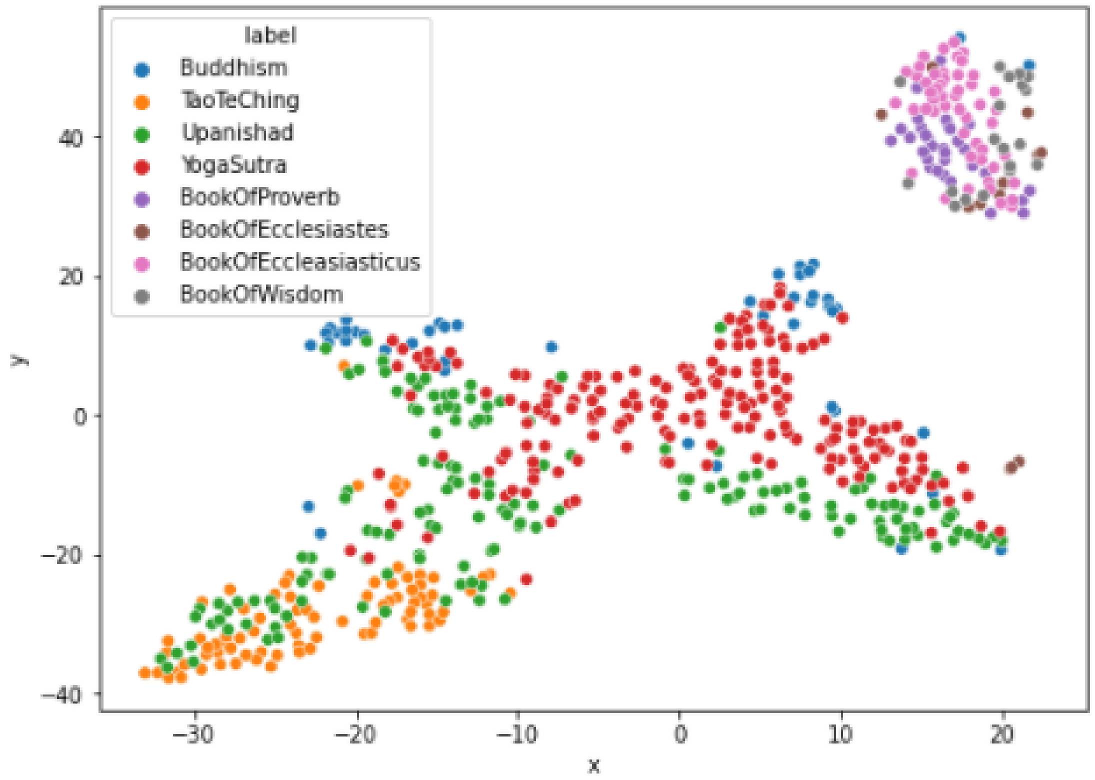


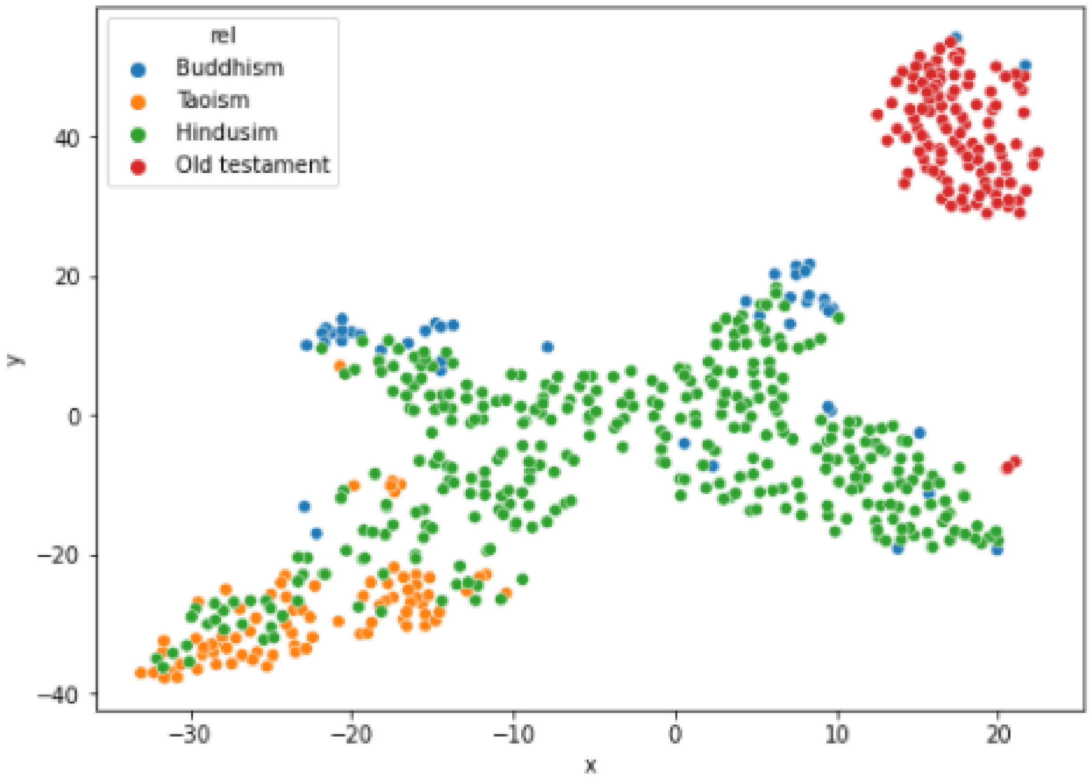
| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.604723 | 0.510290 | 0.495445 | 0.332557 | 0.371286 |
| 3 | ward | 0.585826 | 0.553548 | 0.698478 | 0.457913 | 0.627631 |
| 4 | ward | 0.511152 | 0.607908 | 0.702564 | 0.436472 | 0.639887 |
| 5 | ward | 0.513673 | 0.495200 | 0.703105 | 0.437556 | 0.645189 |
| 2 | complete | 0.817598 | 0.126704 | 0.215867 | 0.001569 | 0.004342 |
| 3 | complete | 0.607592 | 0.380336 | 0.526477 | 0.392522 | 0.447676 |
| 4 | complete | 0.626451 | 0.360984 | 0.645380 | 0.441700 | 0.584799 |
| 5 | complete | 0.569905 | 0.473947 | 0.649501 | 0.431711 | 0.595184 |



| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.604723 | 0.510290 | 0.684429 | 0.466152 | 0.364743 |
| 3 | ward | 0.585826 | 0.553548 | 0.765636 | 0.552127 | 0.571848 |
| 4 | ward | 0.511152 | 0.607908 | 0.767949 | 0.514313 | 0.578370 |
| 5 | ward | 0.513673 | 0.495200 | 0.768490 | 0.516060 | 0.583671 |
| 2 | complete | 0.817598 | 0.126704 | 0.416765 | 0.004490 | 0.004342 |
| 3 | complete | 0.607592 | 0.380336 | 0.722057 | 0.539486 | 0.434639 |
| 4 | complete | 0.626451 | 0.360984 | 0.789025 | 0.546445 | 0.538681 |
| 5 | complete | 0.569905 | 0.473947 | 0.780334 | 0.521176 | 0.539005 |

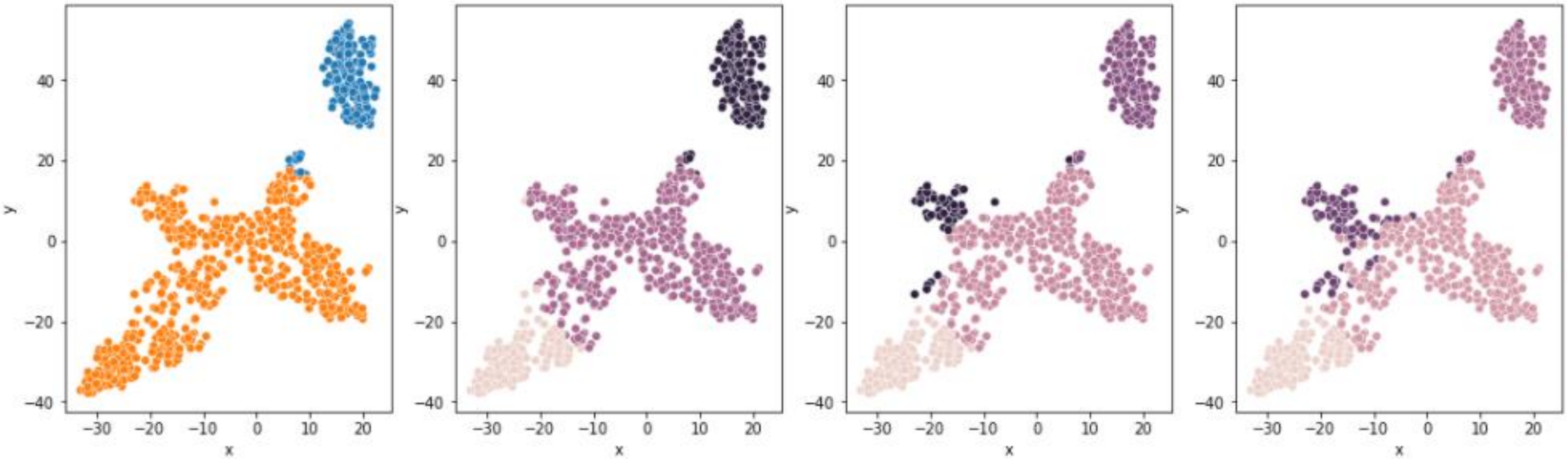
# T-SNE


label
Buddhism
TaoTeChing
Upanishad
YogaSutra
BookOfProverb
BookOfEcclesiastes
BookOfEccleasiasticus
BookOfWisdom
40
20
0
−20
−40
y
−30
−20
−10
0
10
20
x


rel
Buddhism
Taoism
Hindusim
Old testament
40
20
0
−20
−40
y
−30
−20
−10
0
10
20
x

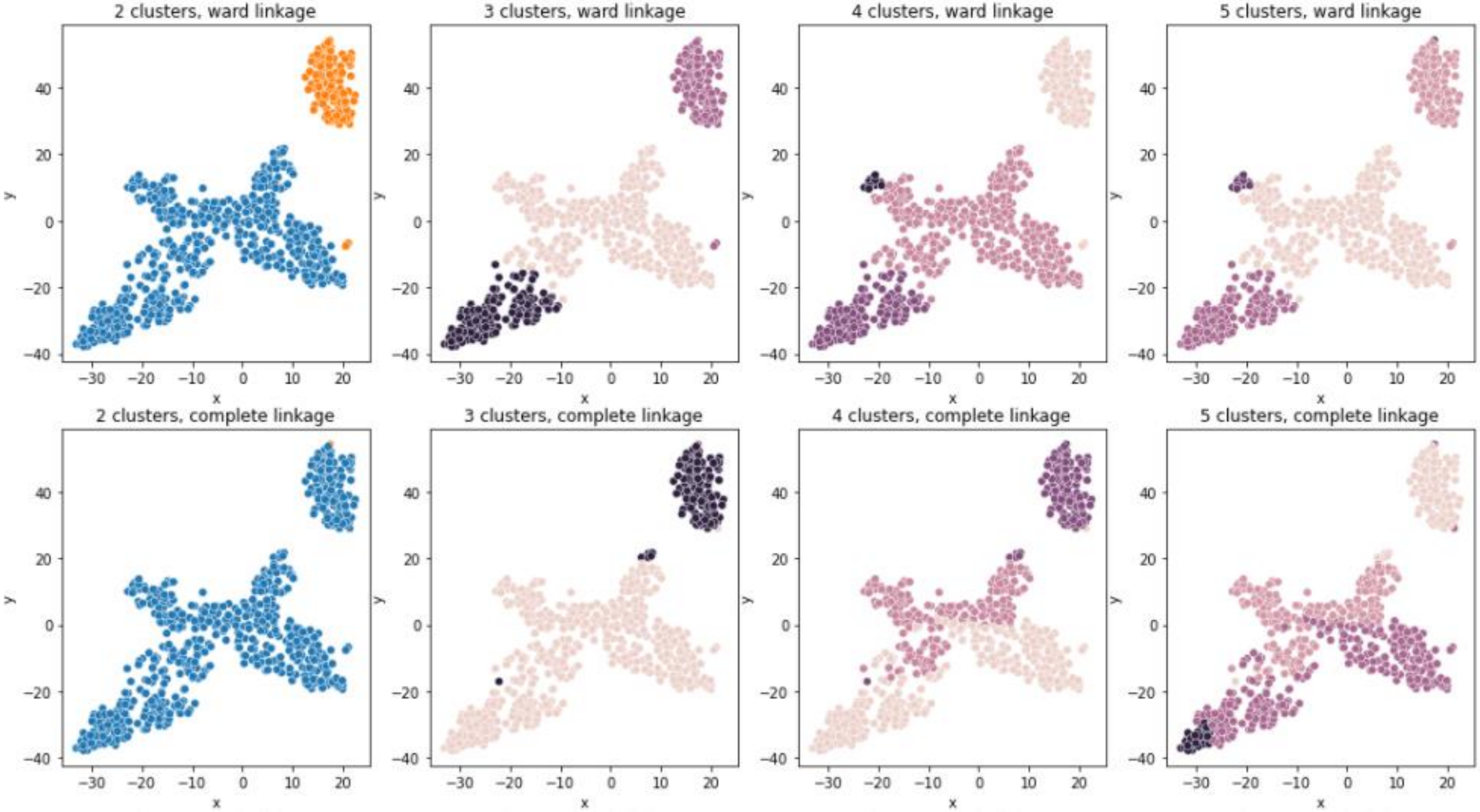
# K-MEANS PO T-SNE

## (BEZ REDUKCJI WYMIARÓW)

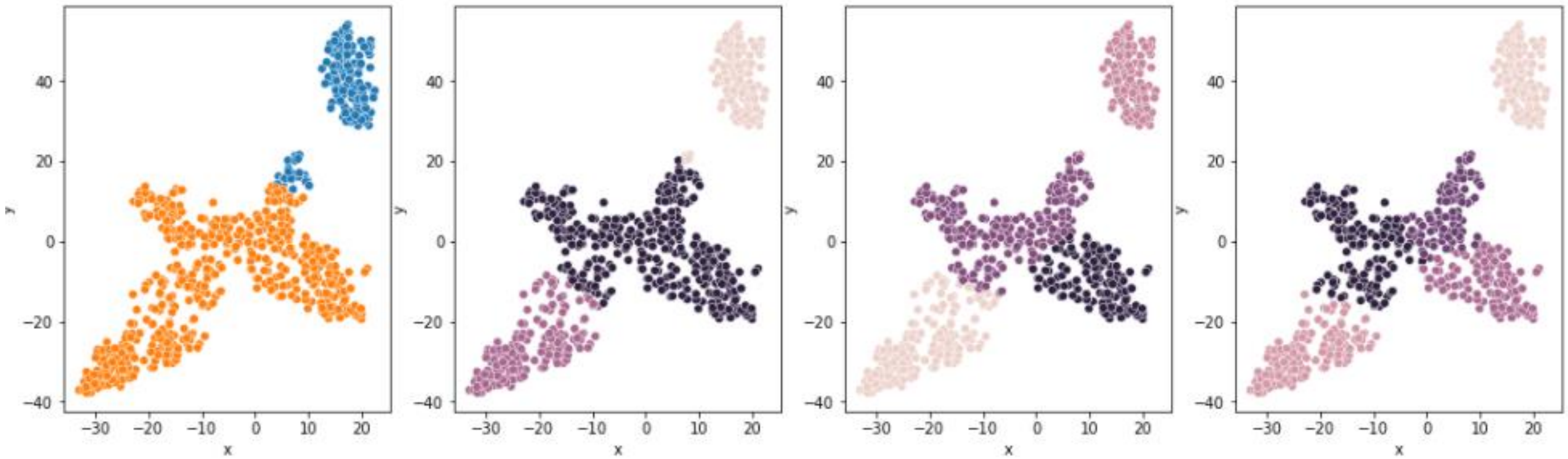


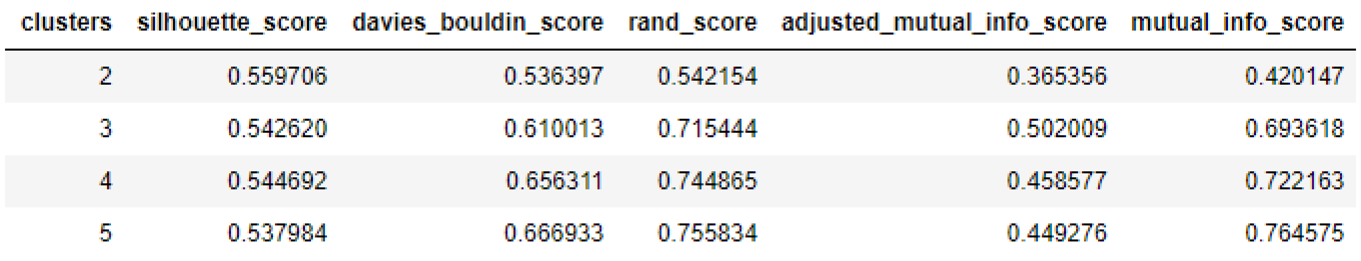
# AGGLOMERATIVE CLUSTERING PO T-SNE

## (BEZ REDUKCJI WYMIARÓW)

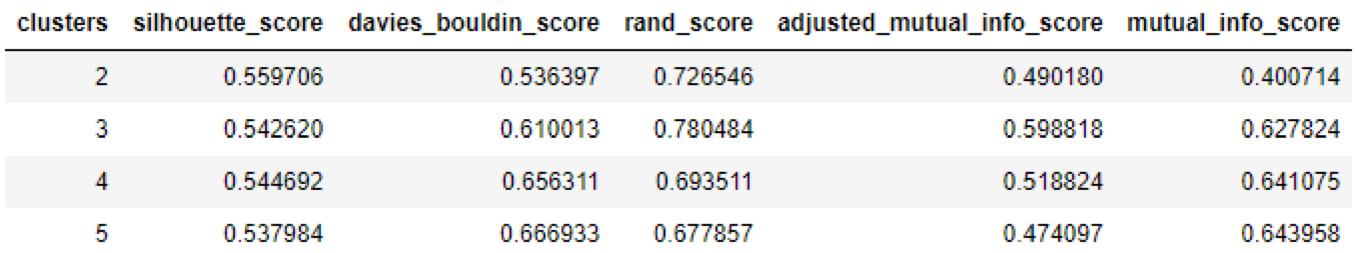


# K-MEANS PO T-SNE



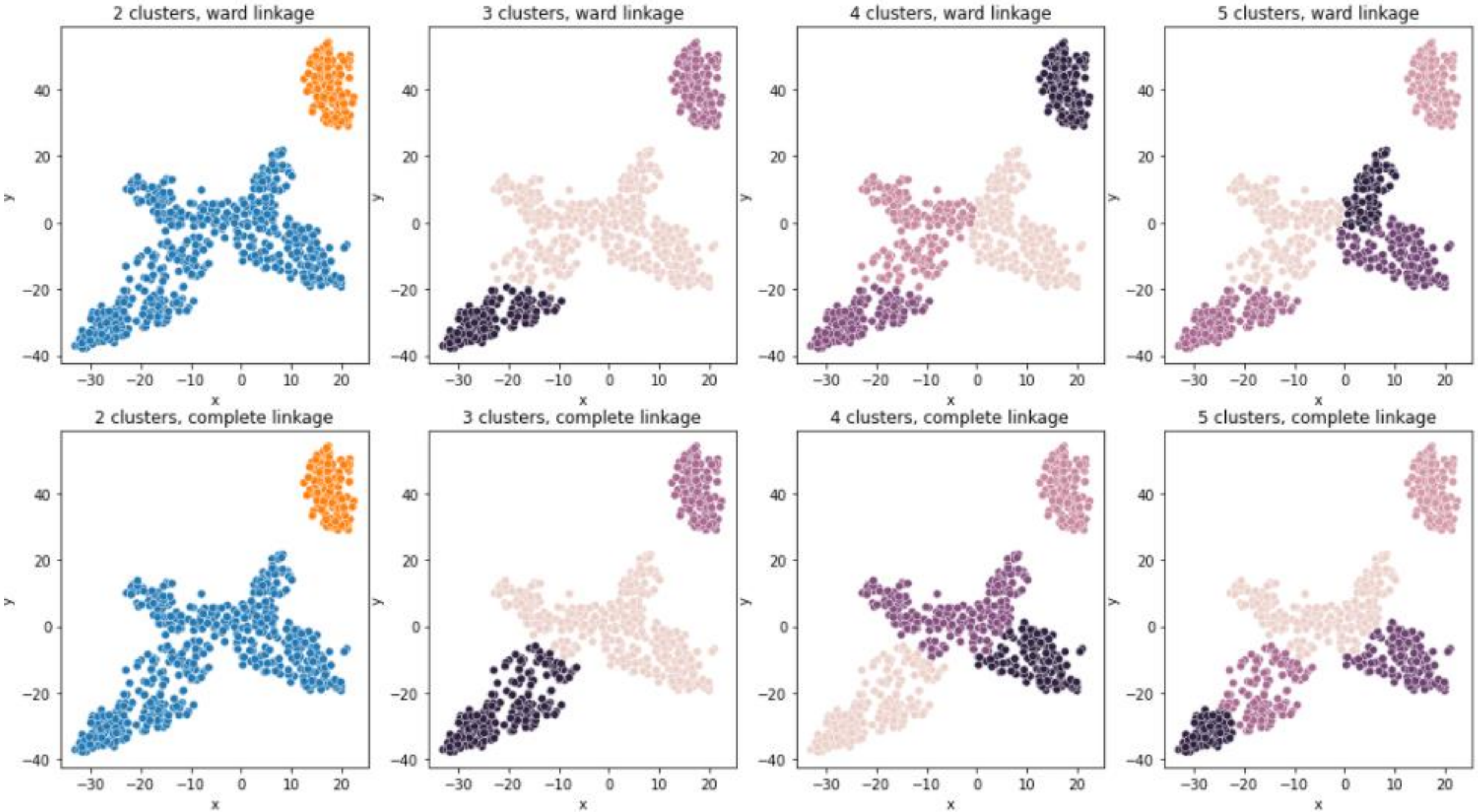


| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.559706 | 0.536397 | 0.542154 | 0.365356 | 0.420147 |
| 3 | 0.542620 | 0.610013 | 0.715444 | 0.502009 | 0.693618 |
| 4 | 0.544692 | 0.656311 | 0.744865 | 0.458577 | 0.722163 |
| 5 | 0.537984 | 0.666933 | 0.755834 | 0.449276 | 0.764575 |

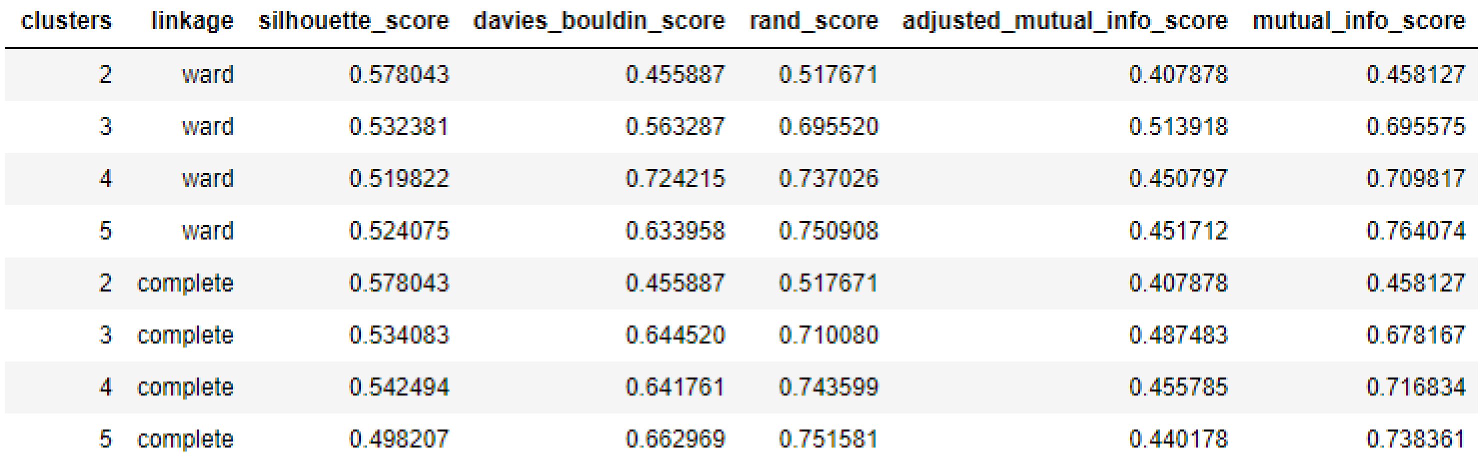


| clusters | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|
| 2 | 0.559706 | 0.536397 | 0.726546 | 0.490180 | 0.400714 |
| 3 | 0.542620 | 0.610013 | 0.780484 | 0.598818 | 0.627824 |
| 4 | 0.544692 | 0.656311 | 0.693511 | 0.518824 | 0.641075 |
| 5 | 0.537984 | 0.666933 | 0.677857 | 0.474097 | 0.643958 |

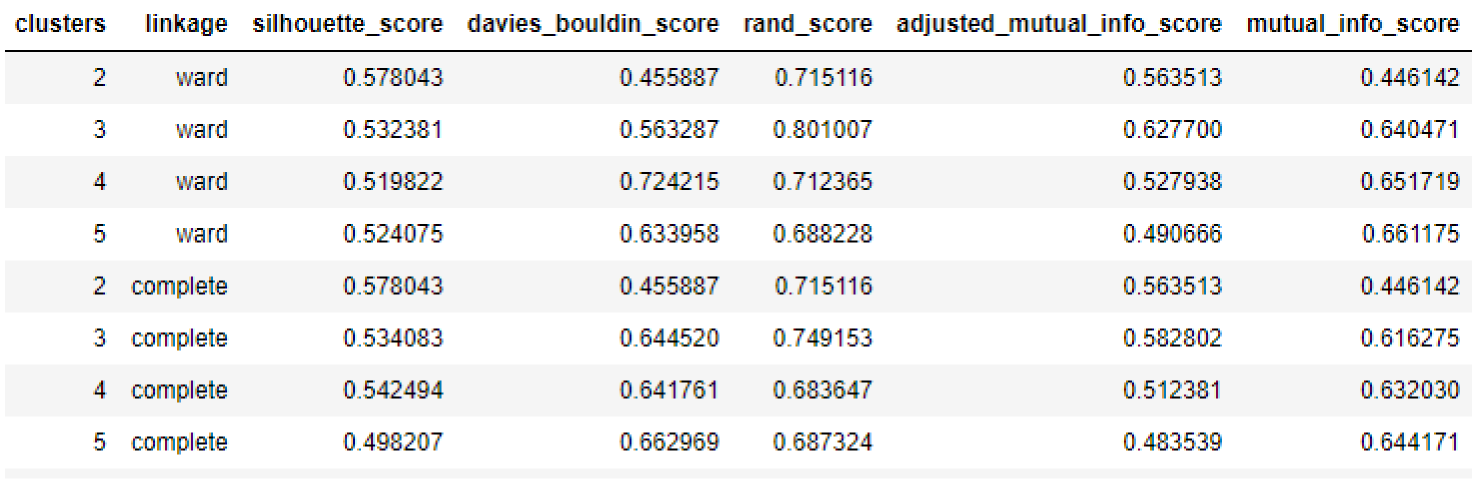
# AGGLOMERATIVE CLUSTERING PO T-SNE



# AGGLOMERATIVE CLUSTERING PO T-SNE



| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.578043 | 0.455887 | 0.517671 | 0.407878 | 0.458127 |
| 3 | ward | 0.532381 | 0.563287 | 0.695520 | 0.513918 | 0.695575 |
| 4 | ward | 0.519822 | 0.724215 | 0.737026 | 0.450797 | 0.709817 |
| 5 | ward | 0.524075 | 0.633958 | 0.750908 | 0.451712 | 0.764074 |
| 2 | complete | 0.578043 | 0.455887 | 0.517671 | 0.407878 | 0.458127 |
| 3 | complete | 0.534083 | 0.644520 | 0.710080 | 0.487483 | 0.678167 |
| 4 | complete | 0.542494 | 0.641761 | 0.743599 | 0.455785 | 0.716834 |
| 5 | complete | 0.498207 | 0.662969 | 0.751581 | 0.440178 | 0.738361 |

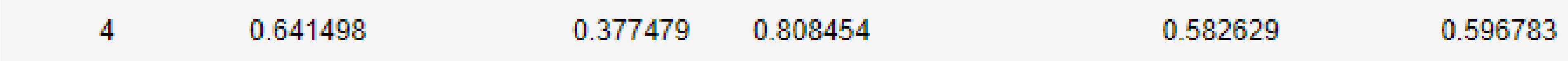


| clusters | linkage | silhouette_score | davies_bouldin_score | rand_score | adjusted_mutual_info_score | mutual_info_score |
|---|---|---|---|---|---|---|
| 2 | ward | 0.578043 | 0.455887 | 0.715116 | 0.563513 | 0.446142 |
| 3 | ward | 0.532381 | 0.563287 | 0.801007 | 0.627700 | 0.640471 |
| 4 | ward | 0.519822 | 0.724215 | 0.712365 | 0.527938 | 0.651719 |
| 5 | ward | 0.524075 | 0.633958 | 0.688228 | 0.490666 | 0.661175 |
| 2 | complete | 0.578043 | 0.455887 | 0.715116 | 0.563513 | 0.446142 |
| 3 | complete | 0.534083 | 0.644520 | 0.749153 | 0.582802 | 0.616275 |
| 4 | complete | 0.542494 | 0.641761 | 0.683647 | 0.512381 | 0.632030 |
| 5 | complete | 0.498207 | 0.662969 | 0.687324 | 0.483539 | 0.644171 |

# POSUMOWANIE

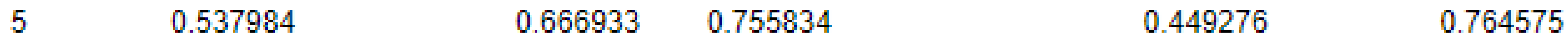
- Bez redukcji wymiarów
  - Najlepiej dla 4 lub 5 klastrów, linkage = complete, sprawdzając z labelami tekstów

- Po PCA
  - Najlepiej dla 4 klastrów Kmeans, sprawdzając labele



| | | | | | |
|---|---|---|---|---|---|
| 4 | 0.641498 | 0.377479 | 0.808454 | 0.582629 | 0.596783 |

- Po T-SNE
  - Najlepije dla 5 klastrów Kmeans sprawdzając labele



| | | | | | |
|---|---|---|---|---|---|
| 5 | 0.537984 | 0.666933 | 0.755834 | 0.449276 | 0.764575 |