PROJEKT 2
MILESTONE3

PRZEMYSŁAW OLENDER, DOMINIK PAWLAK

EDA

- Usunięcie skrótowców (brak)
- Lematyzacja
- Usunięcie 'stopwords'
- Usunięcie słów bardzo długich i bardzo krótkich
- Stworzenie ramki statystycznej i przeskalowanie

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

PRZYGOTOWANIE DANYCH - TF IDF, PRZESKALOWANIE

- Stworzyliśmy również ramkę z wykorzystaniem narzędzia TF DIF

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

- 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

ALGORYTMY KLASTROWANIA

- KMeans
- Agglomerative Clustering: linkage ‚single', ‚complete', ‚ward'
- GMM (Gaussian Mixture Models): covariance: ‚full', ‚tied', ‚diag'

METRYKI

- Silhouette score
- Davies bouldin score
- Rand score

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

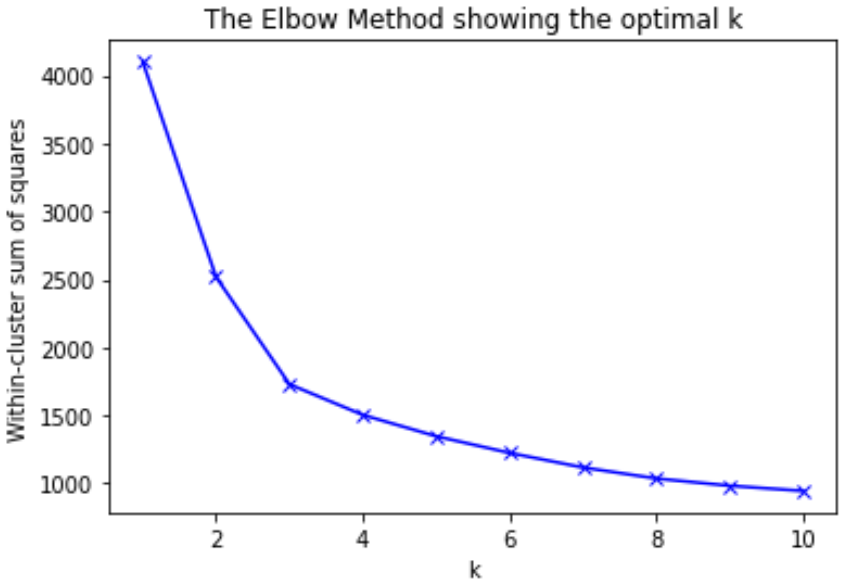
- Adjusted mutual info score

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

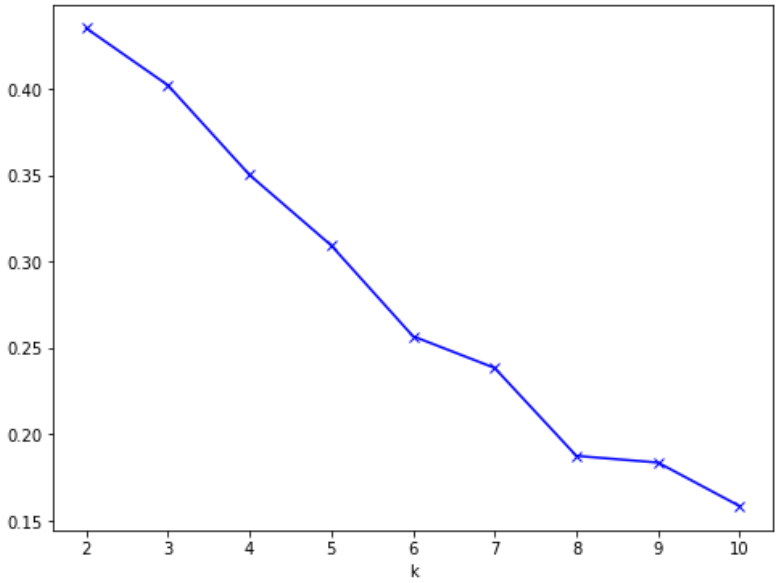
- Mutual info score

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

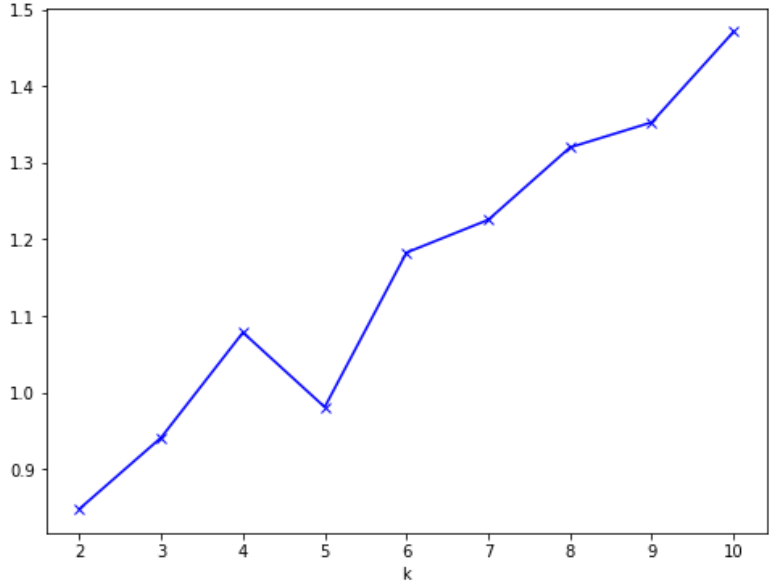
WYBÓR LICZBY KLASTRÓW - NA PEŁNYM ZBIORZE



Wykres łokciowy



Metryka Silhouette

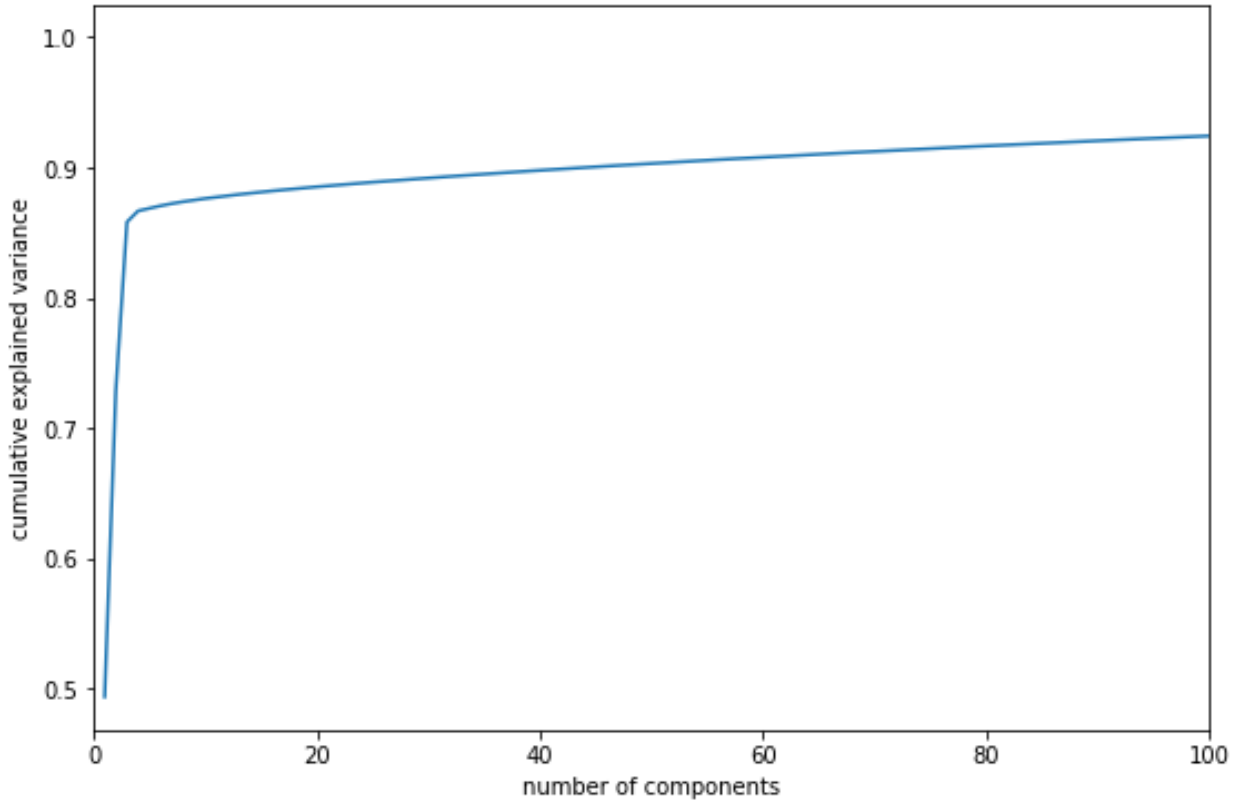


Metryka Davies Bouldin

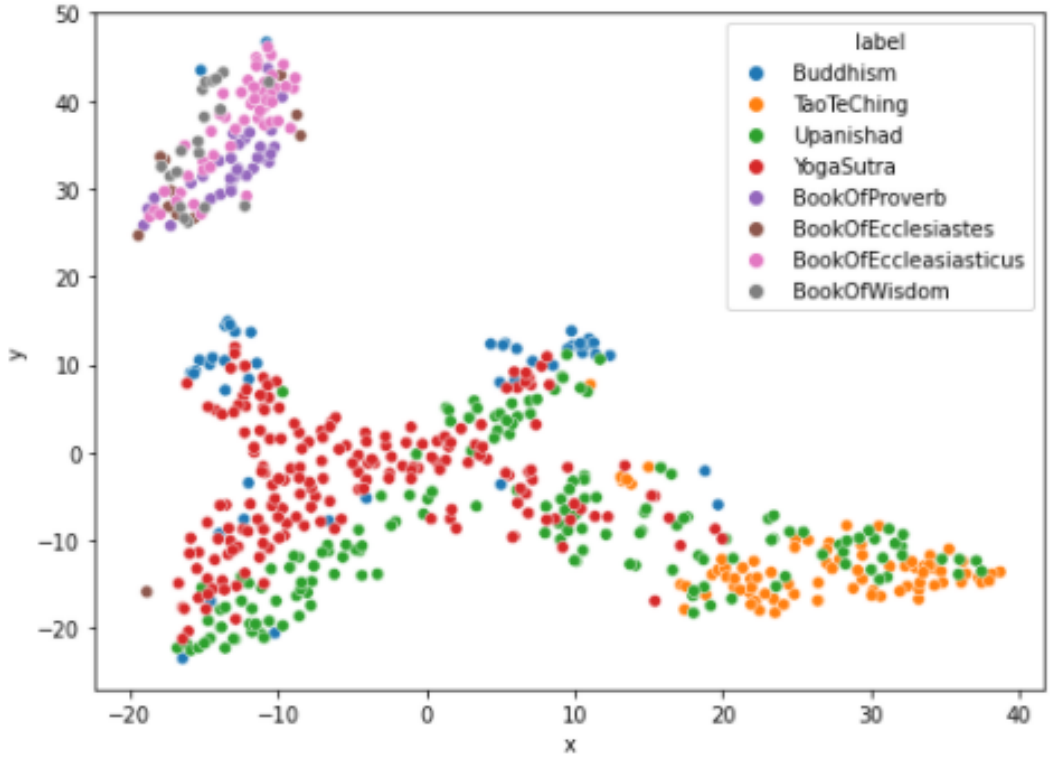
$$DB = \frac{1}{n}\sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)},$$

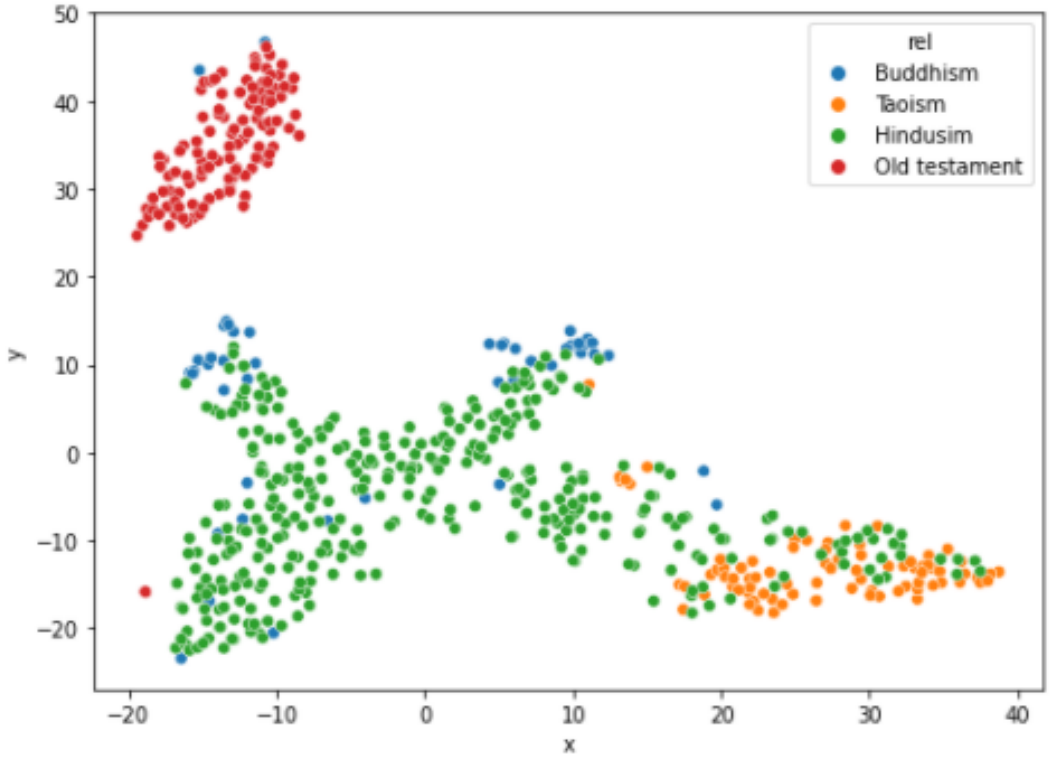
gdzie σ_i jest średnią odległością wszystkich punktów ze skupienia i do jego środka, a $d(c_i, c_j)$ jest odległością pomiędzy środkami skupień i oraz j.

KOMPONENTY PCA

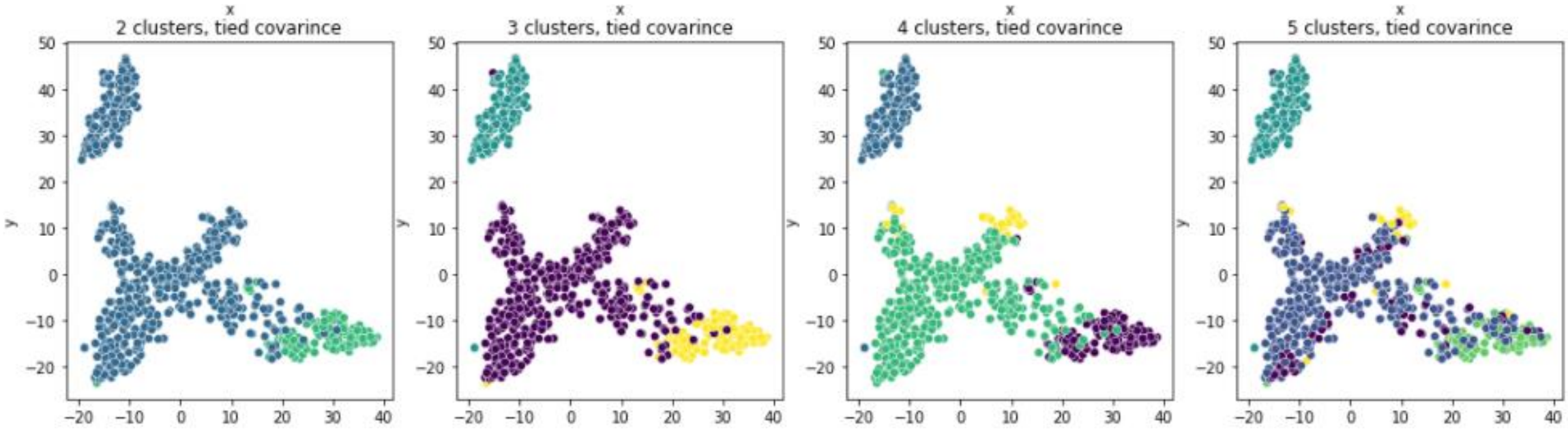


PCA DLA 45 KOMPONENTÓW



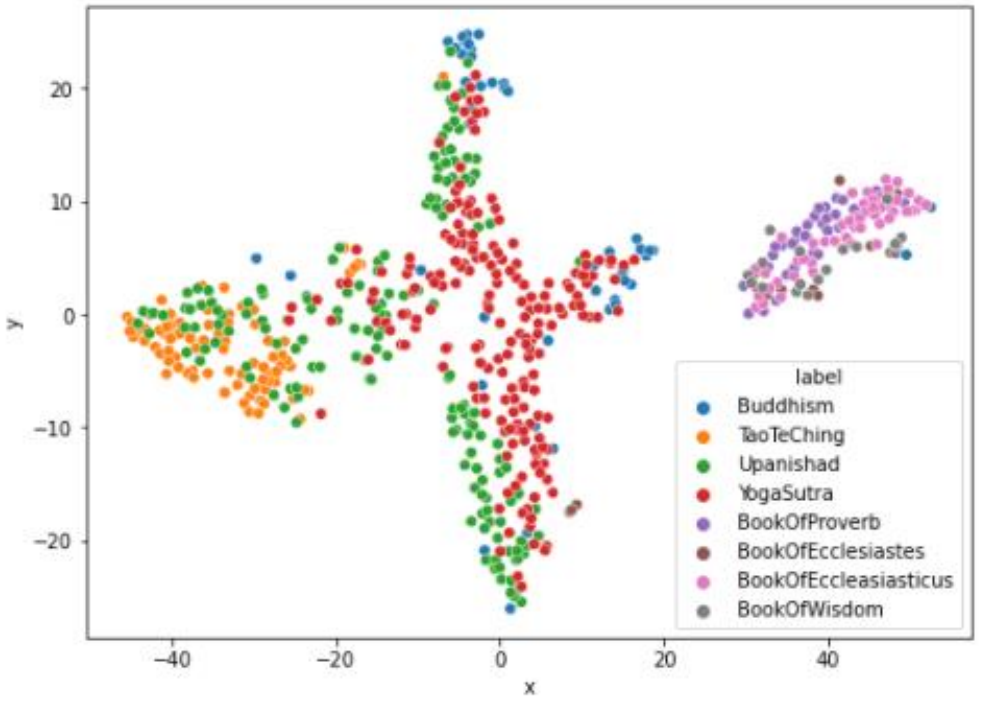


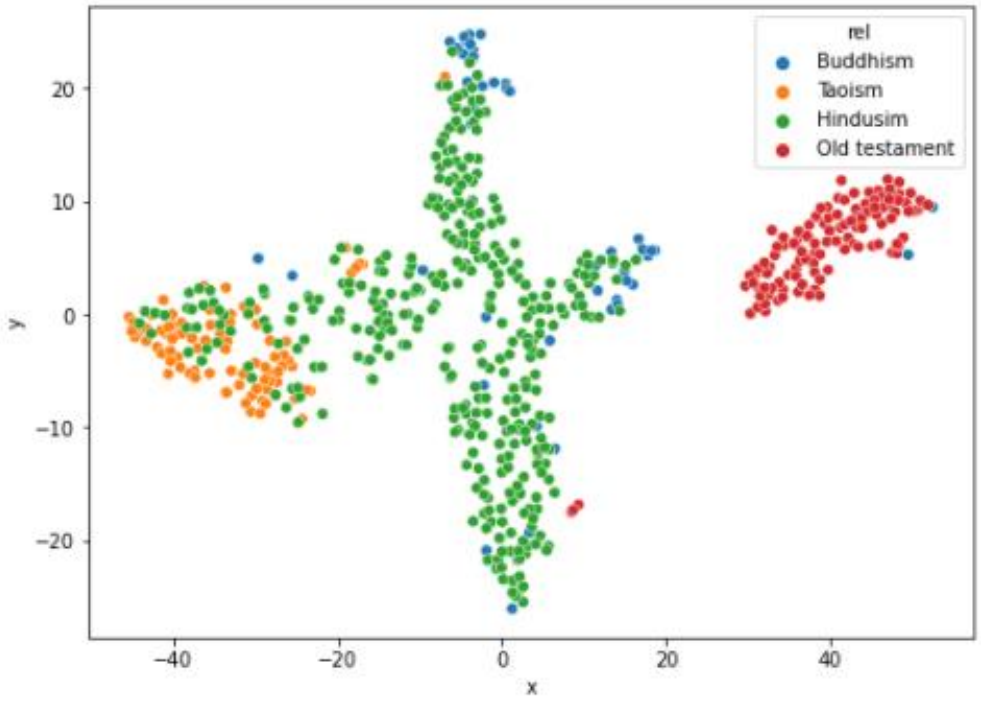
PCA DLA 45 KOMPONENTÓW



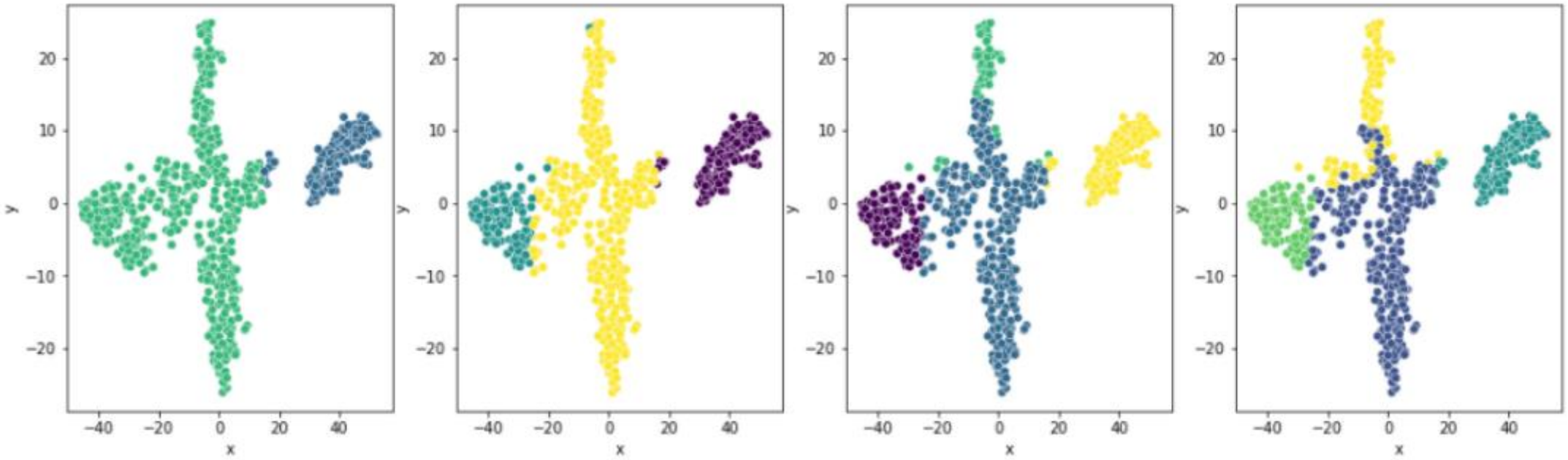
clusters	covariance	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
2	tied	0.319456	1.031972	0.617830	0.391053	0.306398
3	tied	0.434745	0.820805	0.860551	0.741807	0.745419
4	tied	0.412935	0.948276	0.895347	0.767429	0.838767
5	tied	0.182694	2.393333	0.838698	0.731193	0.868887

SPARSE PCA DLA 45 KOMPONENTÓW



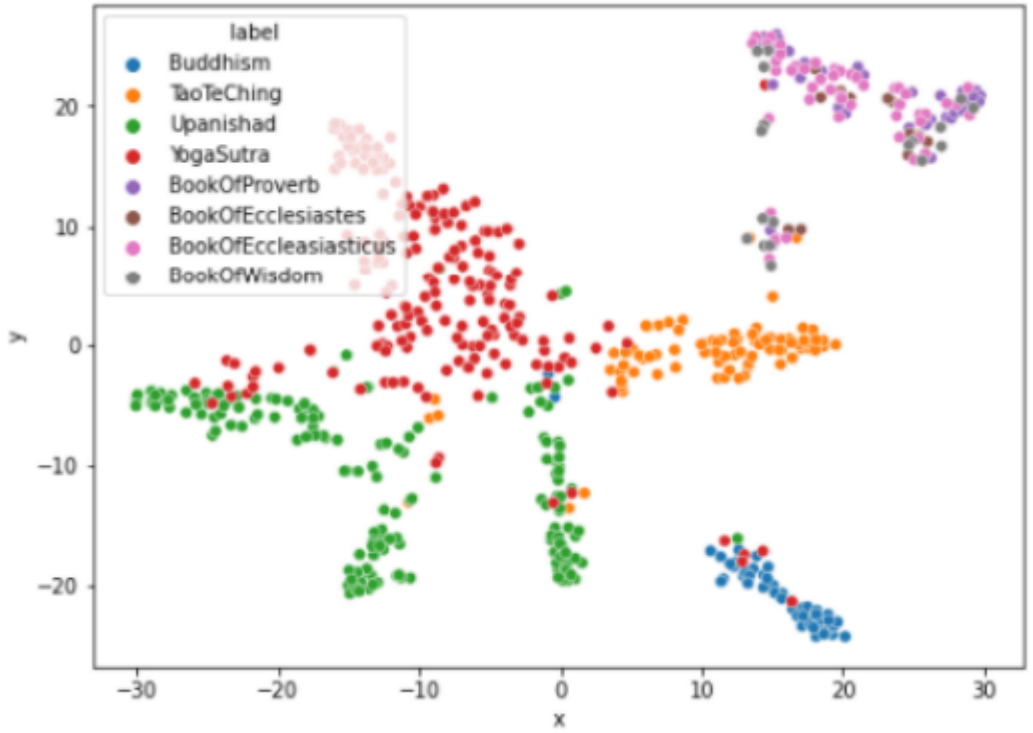


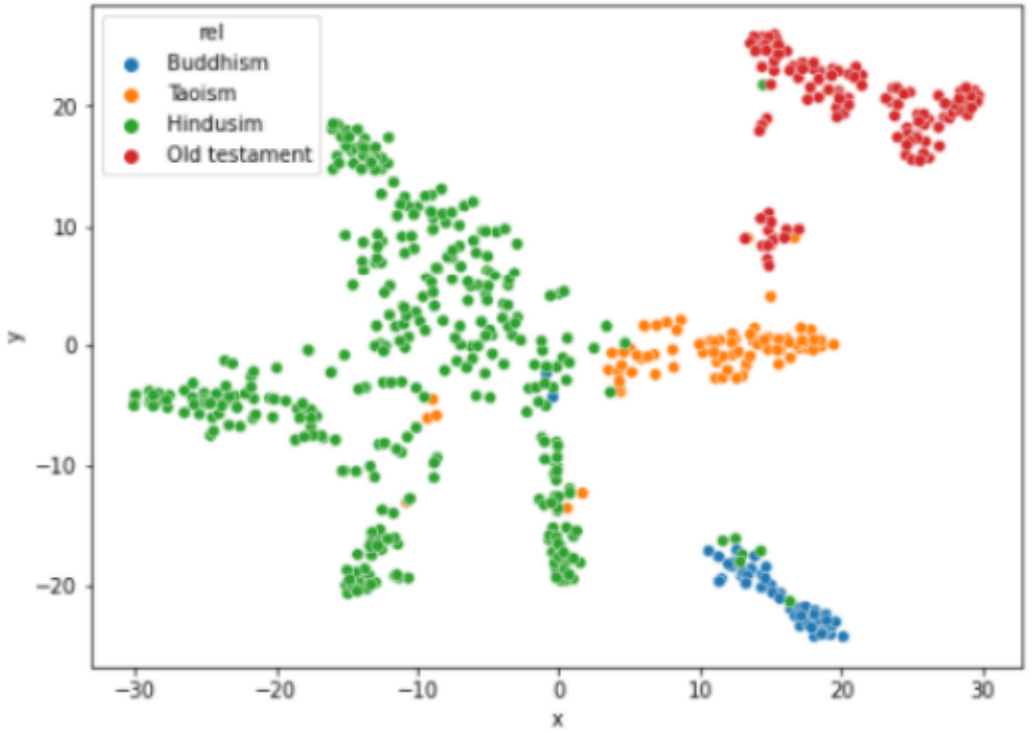
SPARSE PCA DLA 45 KOMPONENTÓW



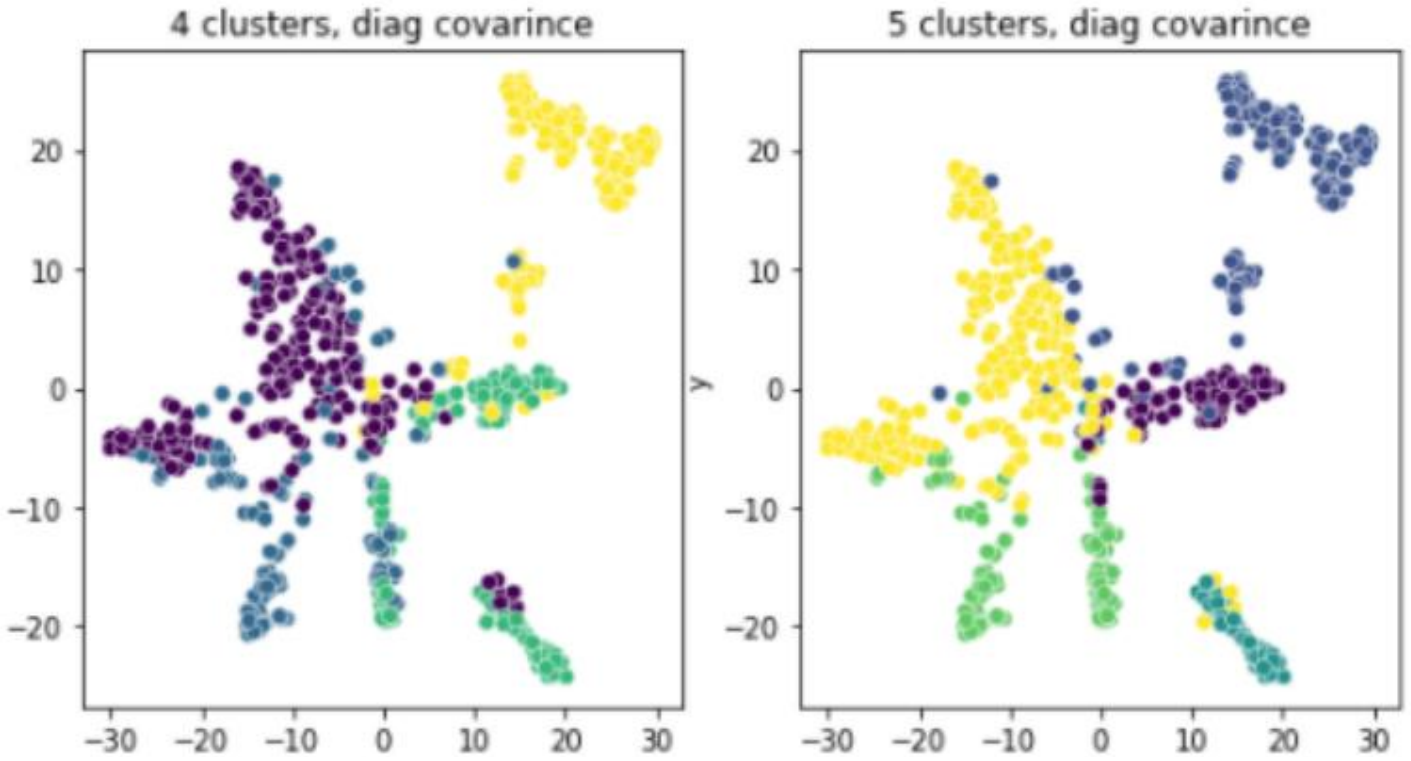
clusters	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
2	0.507076	0.708318	0.724555	0.525175	0.422229
3	0.506940	0.710972	0.807856	0.581368	0.591890
4	0.451944	0.816602	0.802066	0.556585	0.638565
5	0.406508	0.725202	0.768271	0.527087	0.630927

NMF DLA 8 KOMPONENTÓW



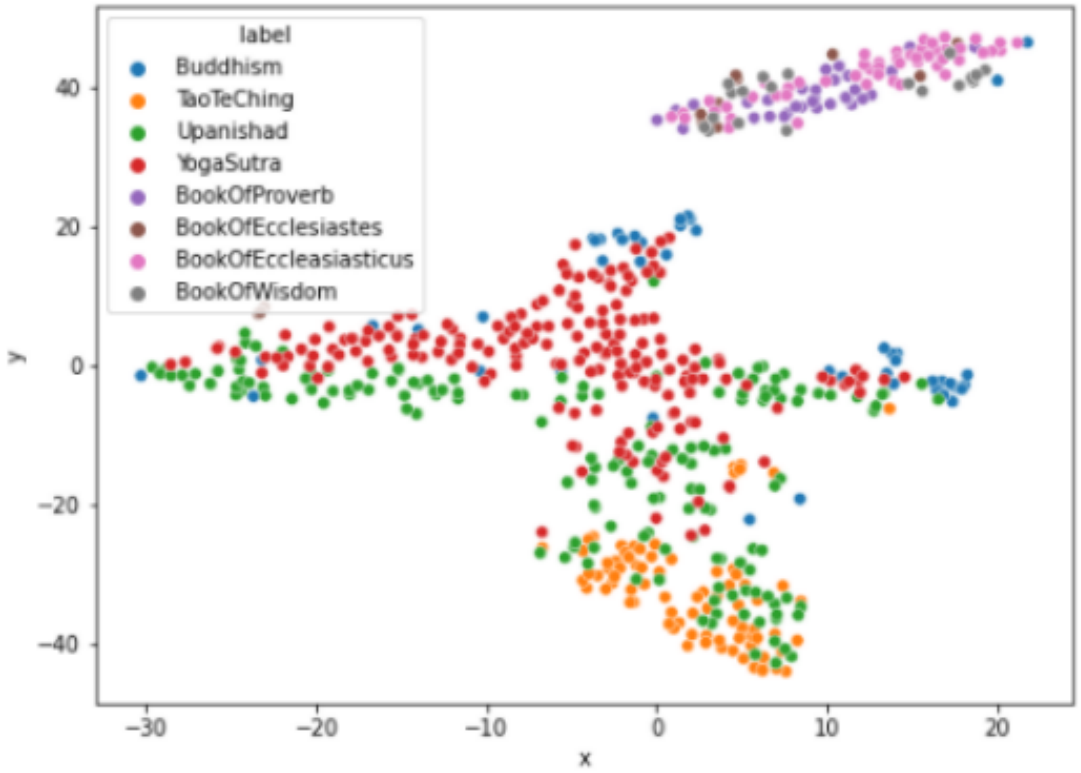


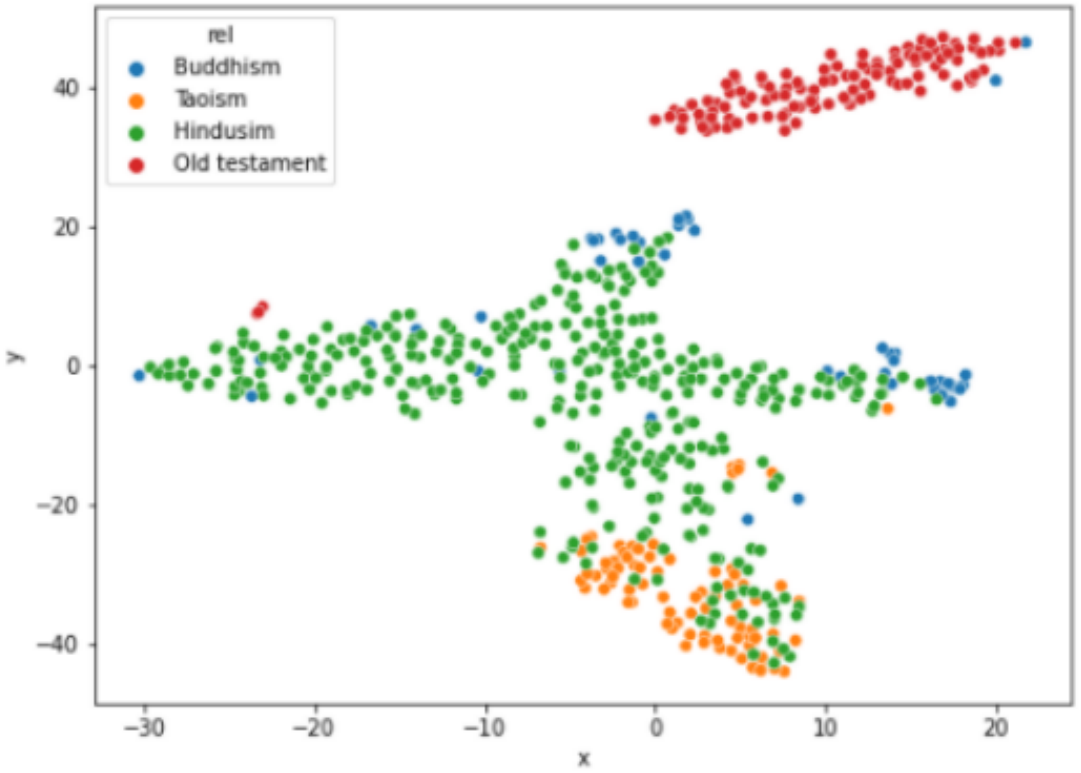
NMF DLA 8 KOMPONENTÓW



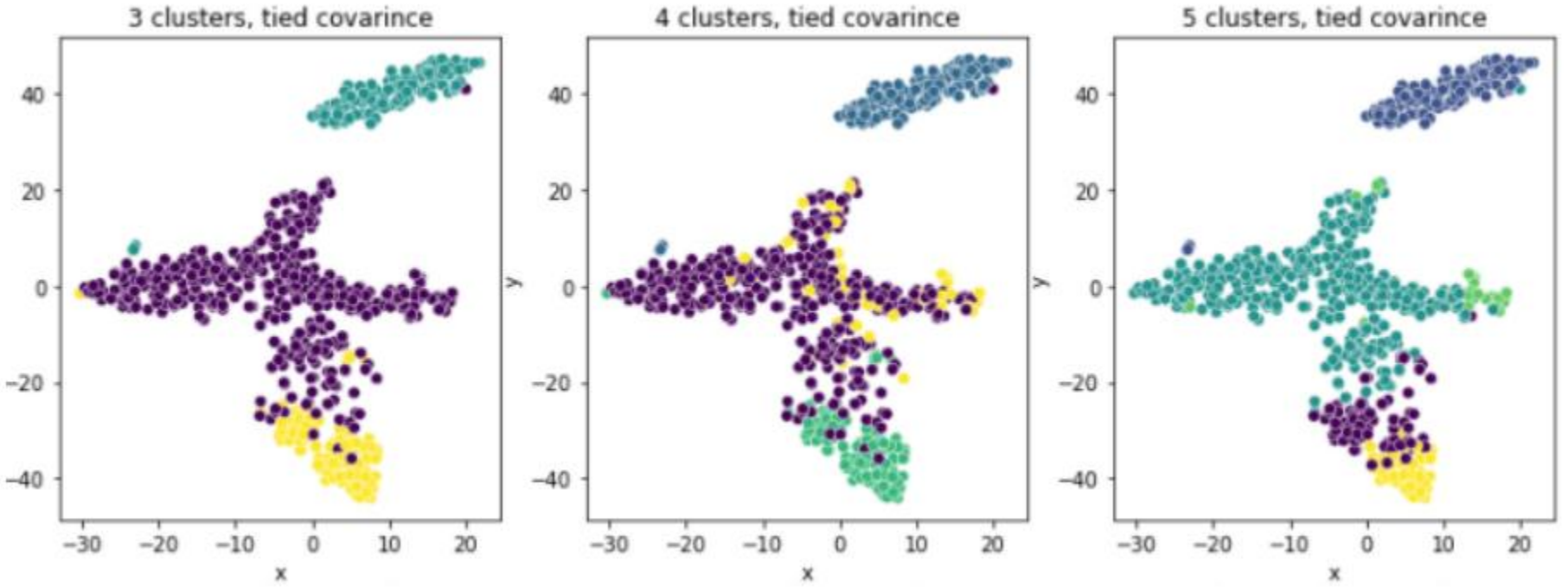
clusters	covariance	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
4	diag	0.176353	1.989555	0.744358	0.527929	0.648687
5	diag	0.210749	1.409956	0.775385	0.616757	0.796627

TRUNCATED SVD DLA 50 KOMPONENTÓW





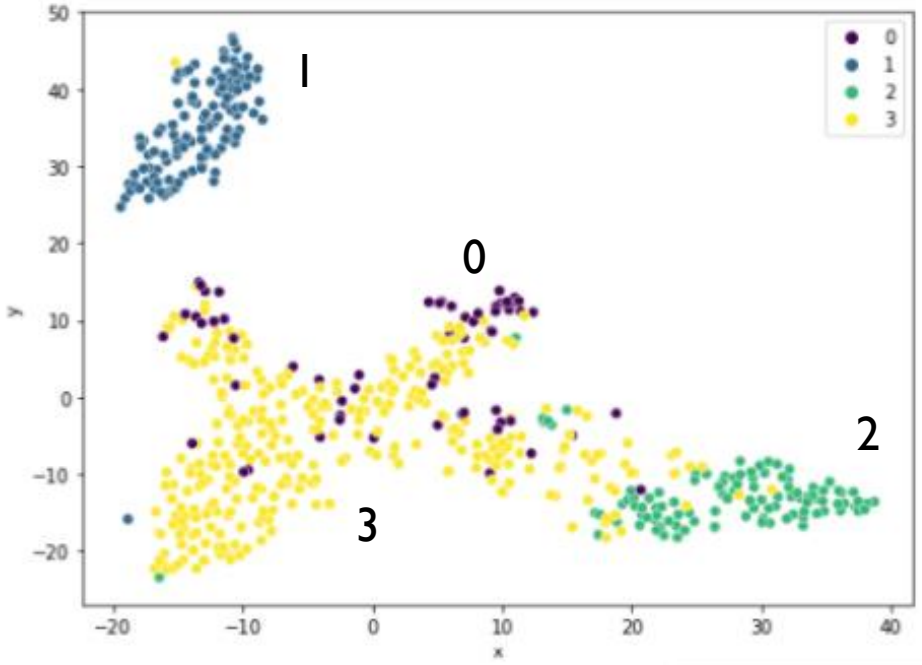
TRUNCATED SVD DLA 50 KOMPONENTÓW

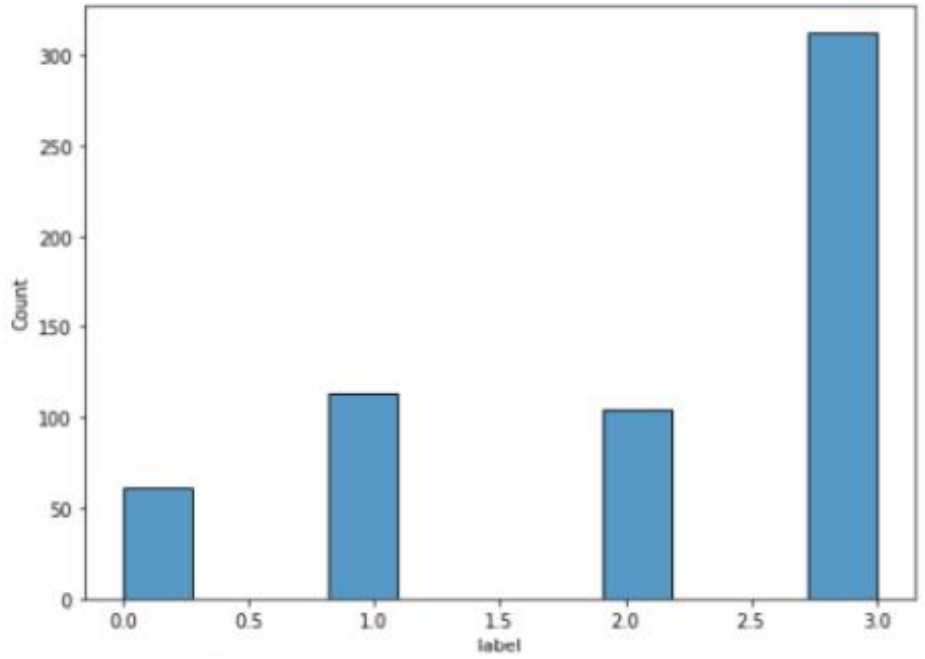


clusters	covariance	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
3	tied	0.432942	0.825195	0.860551	0.741807	0.745419
4	tied	0.281062	1.504503	0.840902	0.680186	0.768905
5	tied	0.349039	1.082496	0.833559	0.626290	0.746307

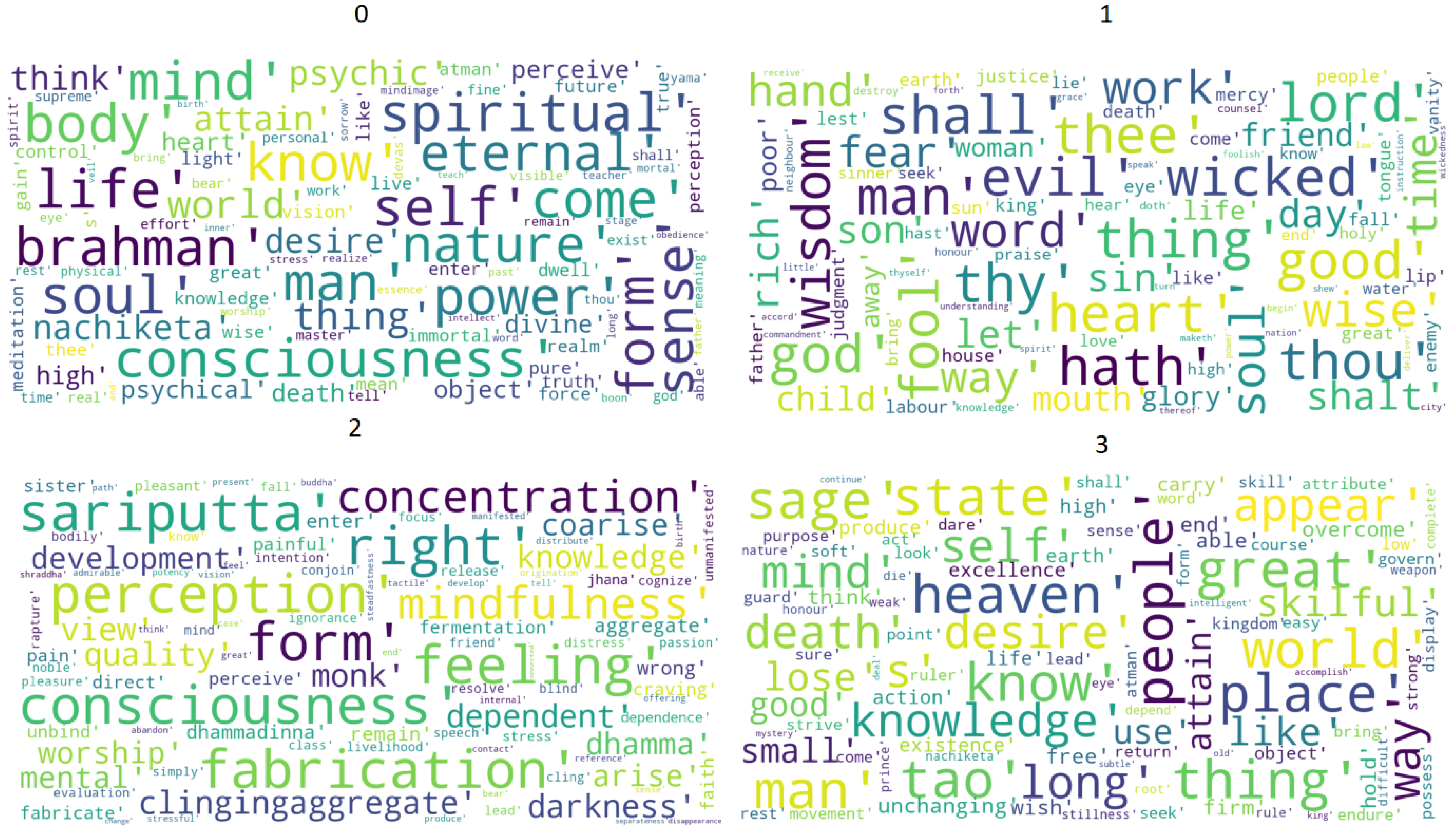
NAJLEPSZE KLASTROWANIA

GMM DLA 4 KLASTRÓW Z COVARIANCJĄ 'TIED', NA ZBIORZE ZREDUKOWANYM PRZY POMOCY PCA Z 45 KOMPONENTAMI.

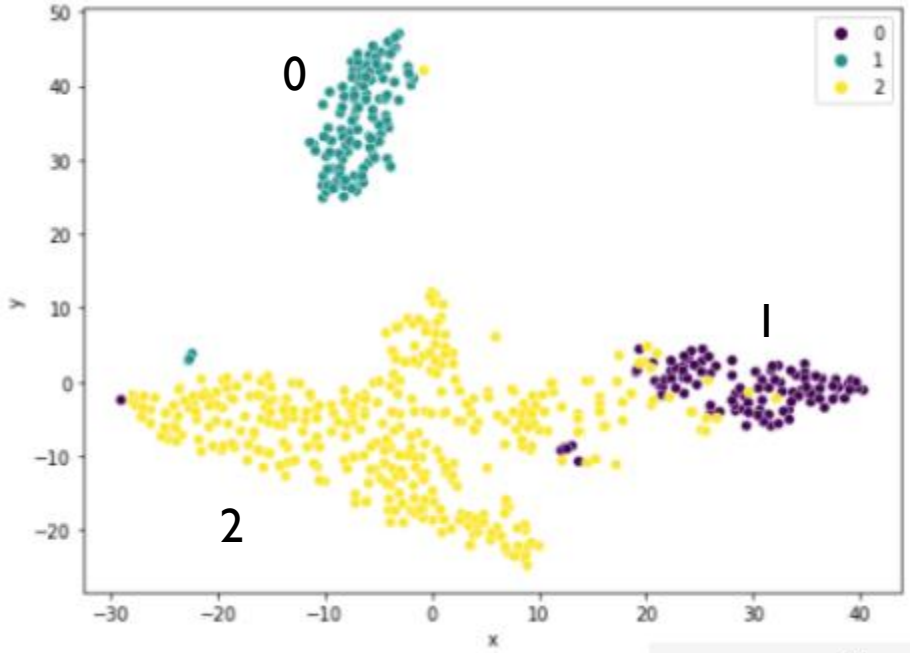


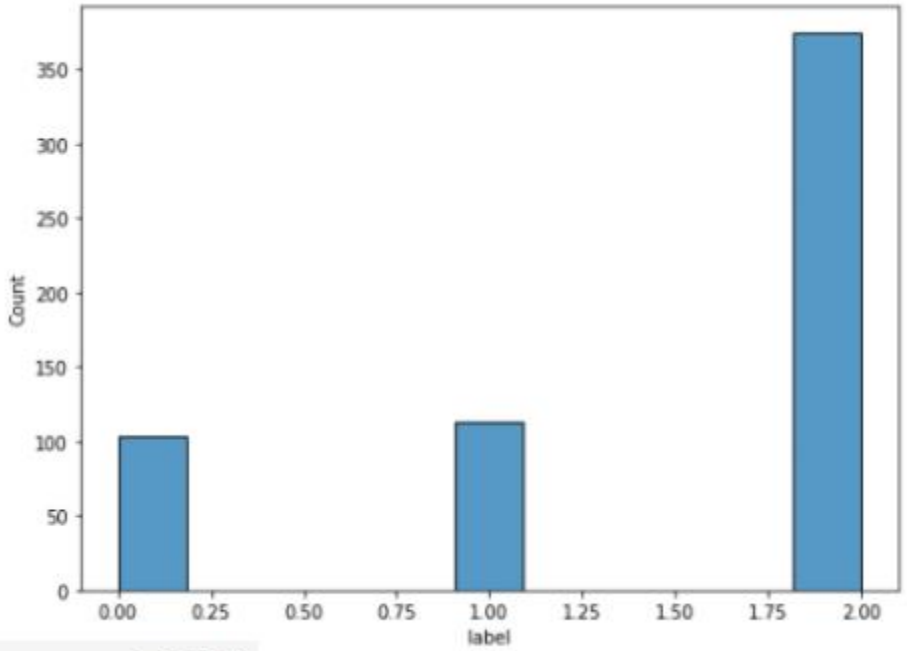


silhouette_score	0.412935
davies_bouldin_score	0.948276
rand_score	0.895347
adjusted_mutual_info_score	0.767429
mutual_info_score	0.838767
calinski_harabasz_score	364.309993


0
think' mind' psychic' atman' perceive' supreme' body' attain' spiritual' eternal' heart' personal' light' know' control' life' self' come' world' vision' work' live' brahman' desire' nature' effort' soul' man' power' form' sense' great' knowledge' nachiketa' thing' divine' wise' immortal' realm' consciousness' high' psychical' death' object' truth' force' pure' meditation' perception' future' shall' mortal' teacher' exist' stage' remain' enter' dwell' thee' time' real' mean' master' physical' rest' spirit' gain' eye' bear' bring' fine' mindimage' yama' true' sorrow' like' devas' visible' teach' inner' stress' realize' essence' worship' intellect' word' long' thou' tell' boon' god' past' obedience' meaning' father' able' end' veil' birth'
1
hand' shall' work' lord' thee' earth' justice' people' lest' fear' woman' friend' come' death' mercy' counsel' evil' wicked' man' wisdom' poor' rich' son' word' thing' day' time' good' thy' sin' heart' wise' god' fool' let' way' hath' soul' thou' shalt' child' mouth' glory' labour' house' love' high' great' enemy' tongue' vanity' know' life' king' hear' eye' sun' fall' holy' end' lip' water' like' praise' honour' hast' away' bring' judgment' father' sinner' seek' neighbour' receive' destroy' forth' lie' grace' foolish' law' speak' doth' instruction' wickedness' little' thyself' understanding' turn' shew' begin' accord' commandment' nation' maketh' spirit' power' deliver' knowledge' thereof' city'
2
sister' path' pleasant' present' fall' buddha' sariputta' concentration' coarise' enter' focus' manifested' bodily' know' development' painful' intention' right' knowledge' distribute' unmanifested' birth' release' conjoin' shraddha' admirable' potency' vision' origination' jhana' cognize' tactile' develop' tell' feel' perception' mindfulness' rapture' ignorance' steadfastness' view' think' mind' case' form' fermentation' aggregate' quality' friend' distress' passion' pain' noble' great' end' feeling' wrong' pleasure' direct' perceive' monk' craving' resolve' blind' consciousness' internal' offering' dependent' dependence' unbind' abandon' dhammadinna' remain' speech' stress' worship' class' livelihood' contact' dhamma' mental' simply' fabrication' reference' cling' arise' faith' evaluation' clingingaggregate' bear' sense' fabricate' lead' darkness' change' stressful' produce' separateness' disappearance'
3
continue' sage' state' shall' carry' skill' attribute' word' appear' high' produce' dare' sense' end' purpose' act' self' overcome' able' course' nature' soft' look' earth' low' complete' mind' great' govern' weapon' excellence' die' people' form' skilful' heaven' guard' think' weak' intelligent' honour' kingdom' easy' death' point' desire' world' display' sure' life' lead' lose' s' ruler' know' eye' atman' attain' strong' accomplish' good' action' place' deal' depend' strive' knowledge' use' like' bring' mystery' existence' small' come' prince' nachiketa' free' return' old' object' subtle' hold' difficult' way' man' tao' long' thing' root' possess' rest' movement' unchanging' wish' stillness' seek' firm' rule' king' endure'

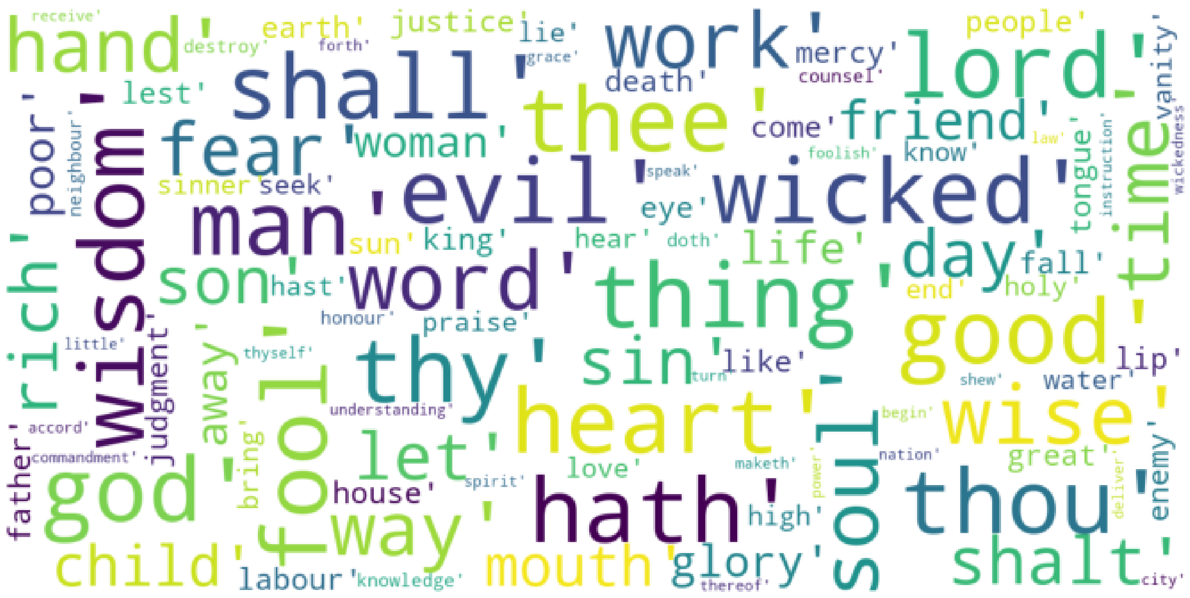
GMM DLA 3 KLASTRÓW Z COVARIANCJĄ 'TIED', NA ZBIORZE ZREDUKOWANYM PRZY POMOCY TRUNCATEDSVD Z 50 KOMPONENTAMI





silhouette_score	0.432961
davies_bouldin_score	0.825250
rand_score	0.860551
adjusted_mutual_info_score	0.741807
mutual_info_score	0.745419
calinski_harabasz_score	442.245899

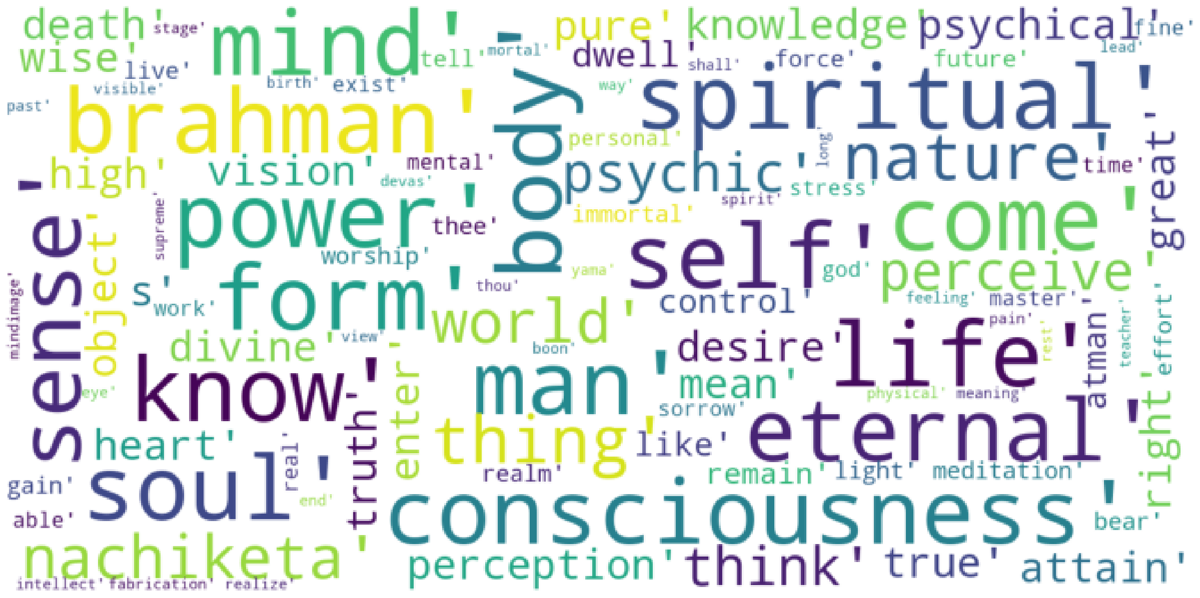
0


hand'
shall'
work'
lord'
thee'
fear'
friend'
woman'
evil'
wicked'
man'
wisdom'
time'
rich'
word'
thing'
day'
son'
good'
thy'
sin'
heart'
wise'
fool'
let'
god'
way'
hath'
soul'
thou'
child'
mouth'
glory'
shalt'
earth'
justice'
people'
lie'
mercy'
death'
come'
lest'
poor'
know'
seek'
sinner'
tongue'
vanity'
eye'
king'
sun'
hear'
life'
fall'
hast'
end'
holy'
honour'
praise'
like'
lip'
away'
judgment'
water'
great'
love'
house'
enemy'
father'
bring'
high'
labour'

1


sage'
state'
excellence'
nature'
attain'
able'
self'
desire'
unchanging'
long'
mind'
great'
people'
heaven'
existence'
good'
way'
firm'
death'
hold'
world'
s'
know'
appear'
place'
knowledge'
lose'
skilful'
small'
free'
overcome'
man'
tao'
like'
thing'
wish'
think'
life'
use'
end'
high'
action'
purpose'
return'
endure'
sure'
easy'
ruler'
shall'
word'
guard'
soft'
act'
low'
carry'
earth'
form'
govern'
produce'
sense'
seek'
course'
kingdom'
attribute'
point'
look'
bring'
strong'
display'

2


death'
mind'
pure'
knowledge'
psychical'
wise'
dwell'
force'
future'
brahman'
body'
spiritual'
high'
vision'
psychic'
nature'
power'
come'
self'
sense'
form'
perceive'
great'
world'
control'
object'
divine'
desire'
life'
know'
man'
mean'
heart'
eternal'
thing'
like'
soul'
truth'
enter'
consciousness'
nachiketa'
perception'
think'
true'
attain'
remain'
light'
meditation'
right'
atman'
live'
exist'
personal'
mental'
immortal'
thee'
worship'
stress'
time'
effort'
realm'
gain'
able'
bear'