Projekt 2 - raport

Wstęp do Uczenia Maszynowego

Przemysław Olender, Dominik Pawlak

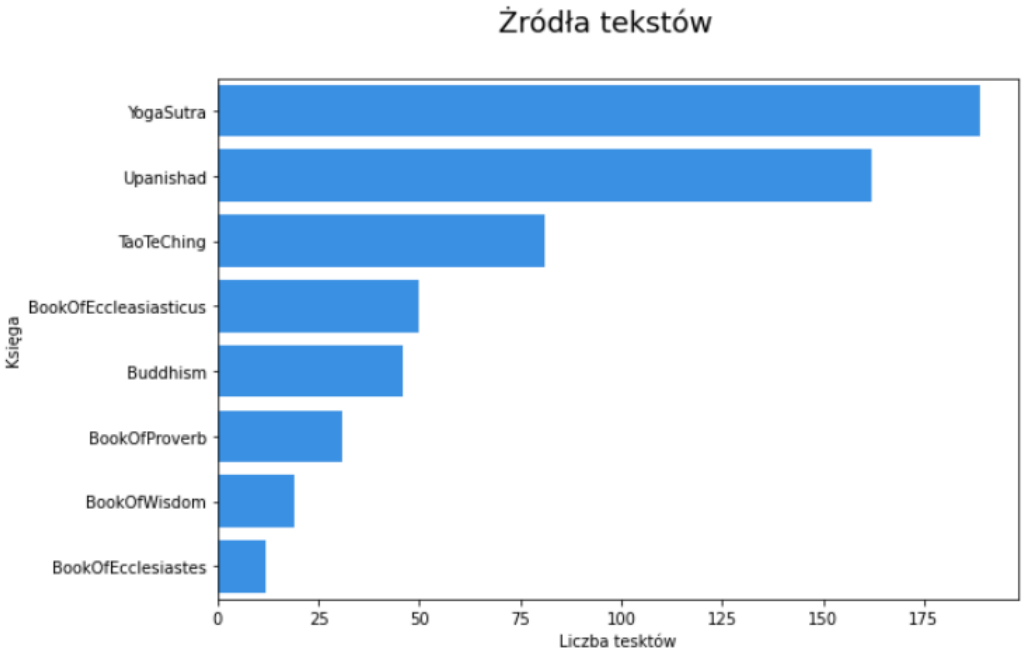
8 czerwca 2021

1 Eksploracyjna analiza danych

Tematem projektu jest klasteryzacja tekstów pochodzących z 8 Świętych ksiąg 4 różnych religii:

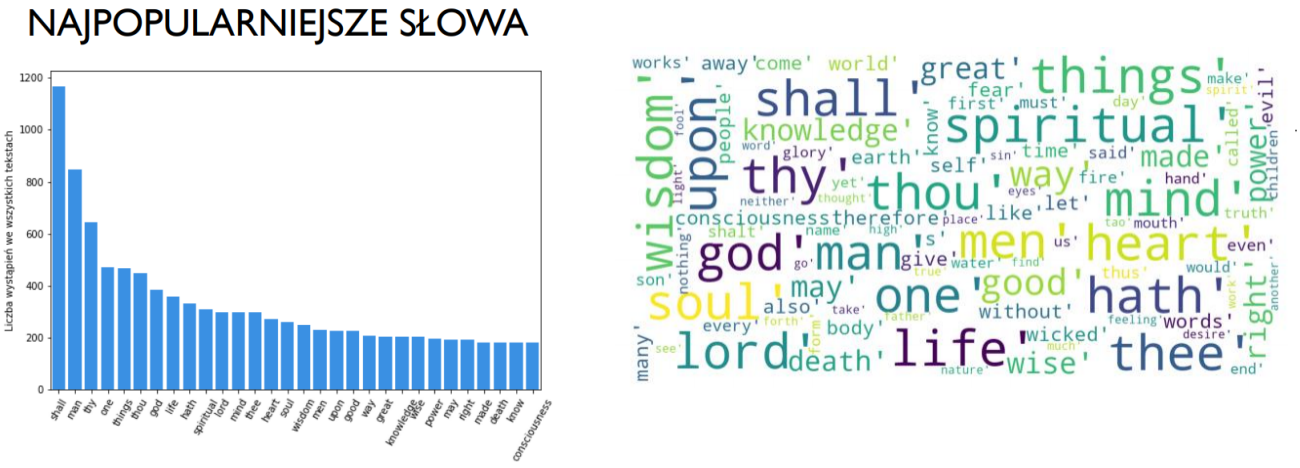
- **Chrześcijaństwo** - Księgi ze Starego Testamentu: Book of Proverbs, Book of Wisdom, Book of Ecclesistes, Book of Eccleasiasticus
- **Hinduizm**: : Yoga Sutra, Upanishad
- **Buddyzm**: jedna księga
- **Taoizm**: jedna księga

Dane zawierają 590 tekstów, w tekstach jest 8267 słów. Utworzono macierz z informcją ile razy dane słowo występuje w tekście. Dostępne są również informacje, z jakiego tekstu pochodzi dany tekst. Najwięcej tekstów pochodzi z Yoga Sutry.

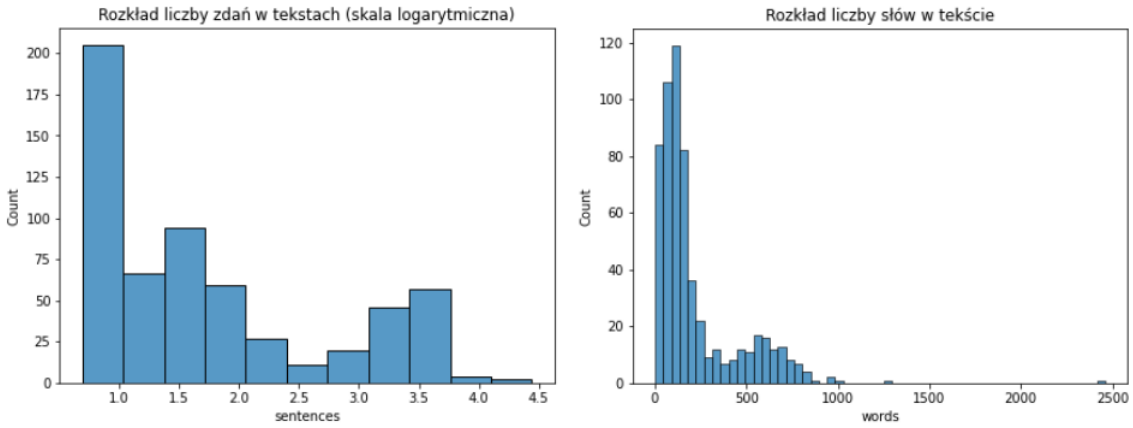


Rysunek 1: Liczba tekstów z danej religii

Najpopularniejszym słowem występującym w tekstach jest 'shall'.

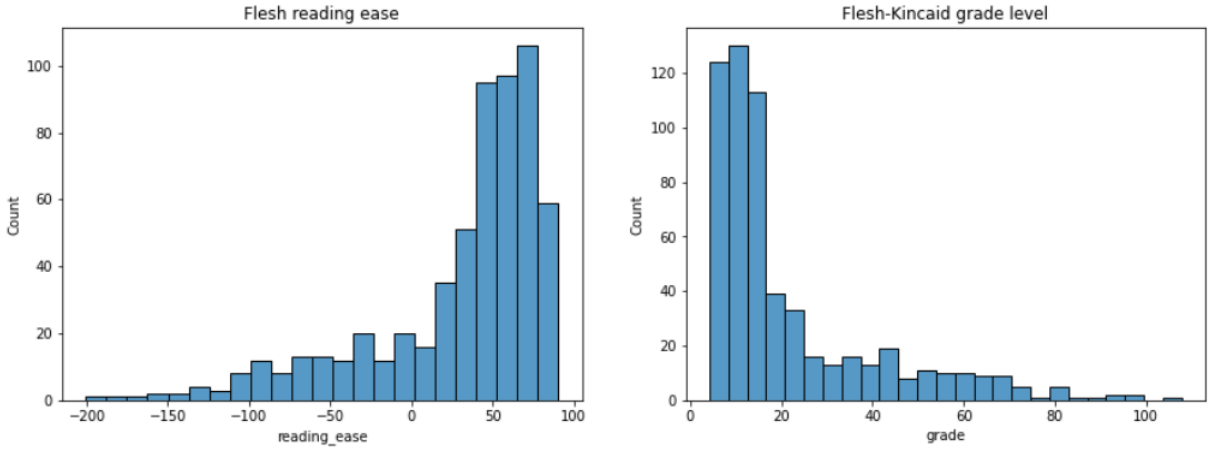


Stworzyliśmy wykresy rozkładów liczby zdań i słów w tekstach. Przeważająca część z nich jest bardzo krótka i zawiera tylko jedno zdanie, co za tym idzie większość tekstów ma poniżej 250 słów.



Rysunek 2: Rozkłady liczby zdań i słów w tekstach

Sprawdziliśmy też jak trudne są teksty, użyliśmy do tego test Flesh reading ease oraz Flesh-Kincaid grade level. Pierwszy z nich jest oblicznay za pomocą wzoru: $206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$. Im większy wynik, tym teskt jest łatwiejszy do przeczytania. Jak widać teksty są dość łatwe. Drugi test określa przybliżony poziom edukacji niezbędny do zrozumienia tekstu. Wynik jest obliczany na podstawie wzoru $0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$. Tutaj im wyższy wynik, tym tekst jest trudniejszy.



Rysunek 3: Łatwość czytania tekstów

2 Praca nad danymi

Na początku naszej pracy sprawdziliśmy czy mamy w naszej ramce danych tzw. skrótowce, czyli słówka typu "don't","aren't", isn't" itp. Okazało się, że takich słów w naszej ramce nie ma, przeszliśmy zatem do lematyzacji.
Lematyzacja, czyli wydobycie ze słów ich korzenia / podstawy słowotwórczej okazała się być bardzo dobrym pomysłem. Pozwoliła zredukować rozmiar naszej ramki danych z 8277 kolumn do 6277. Słownik stworzony ze znalezionych słów zawierał 2577 elementów.
Kolejnym etapem było usunięcie z ramki tzw. 'stopwords'. Pozwoliło nam to jeszcze bardziej zredukować rozmiar ramki danych, o 166 kolumn.
Przyjrzeliśmy się też słowom bardzo krótkim i bardzo długim. Okazało się, że jest ich bardzo niewiele. W większość występują max. 10 razy w całym zbiorze danych (590 rekordów), dlatego je również zdecydowaliśmy się usunąć. Ostateczny rozmiar naszej ramki to 6073 kolumn (z początkowych 8266).
Następnie przeskalowaliśmy naszą ramkę danych za pomocą TF IDF. Metoda ta określa wagę słowa, biorąc pod uwagę liczbę wszystkich słów w danym tekście oraz liczbę wystąpień tego słowa we wszystkich tekstach.

Wartość $tfidf$ wyliczamy ze wzoru: $tfidf_{i,j} = tf_{i,j} * idf_i$, gdzie:
$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ oraz $idf_i = log\frac{|D|}{|\{d:t_i \in d\}|}$, gdzie:

- $n_{i,j}$ - liczba wystąpień słowa i w dokumencie j
- $n_{k,j}$ - suma wszystkich słów w dokumencie j
- $| D |$ - liczba wszystkich dokumentów
- $| \{d : t_i \in d\} |$ - liczba dokumentów zawierających przynajmniej jedno wystąpienie danego tekstu

Ustawiliśmy dolny i górny threshold na poziomie 0.2 i 0.9. Po tych działaniach nasza ramka miała już "tylko" 3366 kolumn. Na podstawie pliku z surowymi tekstami stworzyliśmy ramkę statystyczną. Uwzględniliśmy w niej parametry takie jak: długość tekstu, liczba różnych słów, trudność tekstu, oraz liczbę zdań. Wartości w tej ramce przeskalowaliśmy za pomocą Standard Scaler'a.
Obie ramki złączyliśmy ze sobą. W ten sposób otrzymaliśmy ramkę, na której zaczęliśmy modelowanie.

```
2 X.head()
```

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

5 rows × 3372 columns

Rysunek 4: Ostateczny kształt ramki

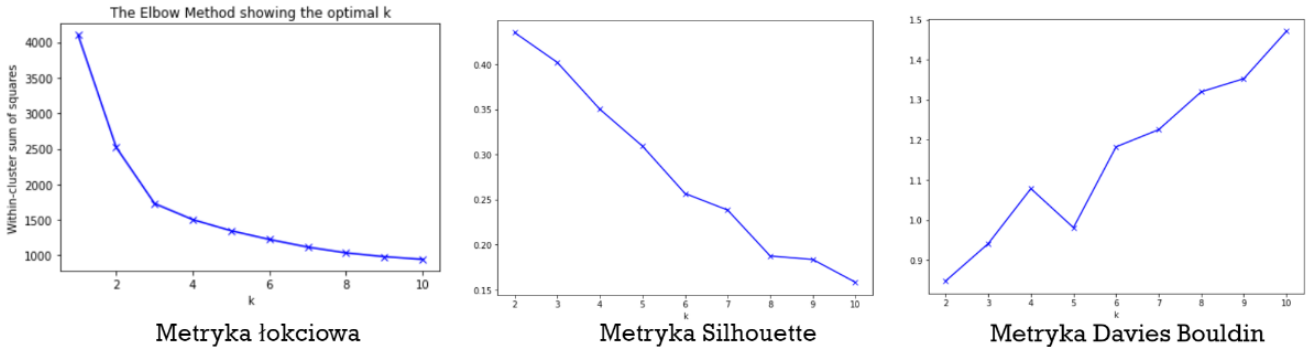
3 Modelowanie

3.1 Algorytmy klastrowania

Używaliśmy 3 algorytmów:

- KMeans
- Agglomerative Clustering
- GMM (Gaussian Mixture Models)

Dla każdego algorytmu sprawdzaliśmy zachowanie dla różnej liczby klastrów wyznaczonej za pomocą trzech metod: Metody Łokcia przy użyciu KMeans (wybór to przegięcie łokcia), Metryki Silhouette'a (im lepszy wynik tym lepiej) oraz metryki Daviesa Bouldina (im mniejszy wynik tym lepiej).



Rysunek 5: Wyznaczanie liczby klastrów

Biorąc pod uwagę wszystkie 3 metody oraz to, że nasze testy pochodzą z 4 różnych religii sprawdzaliśmy klastrowania dla 2-5 klastrów. Nie za to szukaliśmy 8 klastrów, czyli liczby ksiąg.

Dla każdej metody dobieraliśmy hiperparametry; dla Agglomerative Clustering testowaliśmy linkgae: 'single', 'ward' i 'complete', a dla GMM testowaliśmy covarince: 'full', 'tied' i 'diag'.

3.2 Metryki

Do oceny klastrowań używaliśmy następujących metryk, wykorzystaliśmy dostępność labeli tekstów:

- **Silhouette Score** - miara odległości między klastrami, im większa tym dalej od siebie znajdują się klastry. Zakres [-1, 1]
- **Davies-Bouldin Index** - miara podobieństaw między danym klastrem i tym najbardziej podobnym do niego. Podobieństwo to współczynnik odległości wewnątrz klastra. Im mniejszy wynik, tym lepiej. Wzór to:

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

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.

- **Rand Index** - zakłada, że znamy dobre klastry z którymi możemy porównać naszą odpowiedź, sprawdza czy rekordy zostały przyporządkowane to tych samych klastrów. Wzór to:

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

- **Adjusted Rand Index** - podobny do Rand Index, bierze pod uwagę liczbę wszystkich par przypisanych do klastra. Wzór:

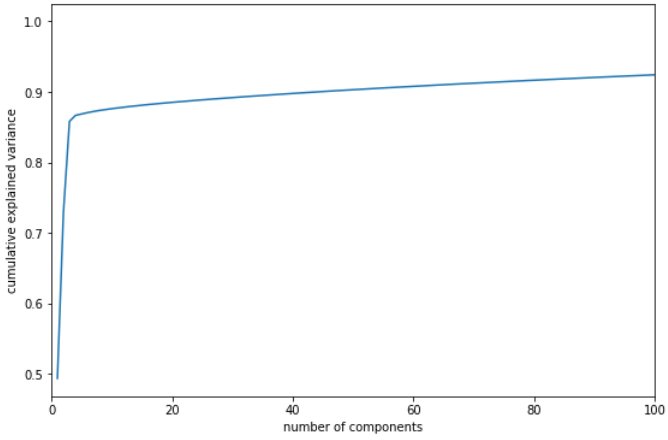
$$ARI = \frac{RI - Expected\ RI}{Max(RI) - Expected\ RI}$$

- **Mutual Information** - mierzy podobieństwo między labelami danych.

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

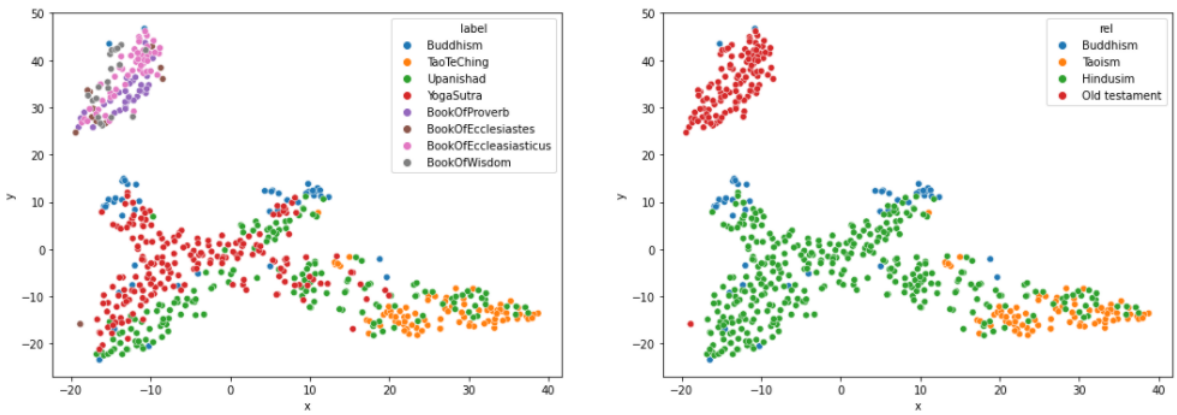
3.3 PCA

Pierwszym algorytmem do redukcji wymiarów, który zastosowaliśmy było PCA. Zaczęliśmy od sprawdzenia ile komponentów wyjaśnia wystarczająco wysoki odsetek wariancji.



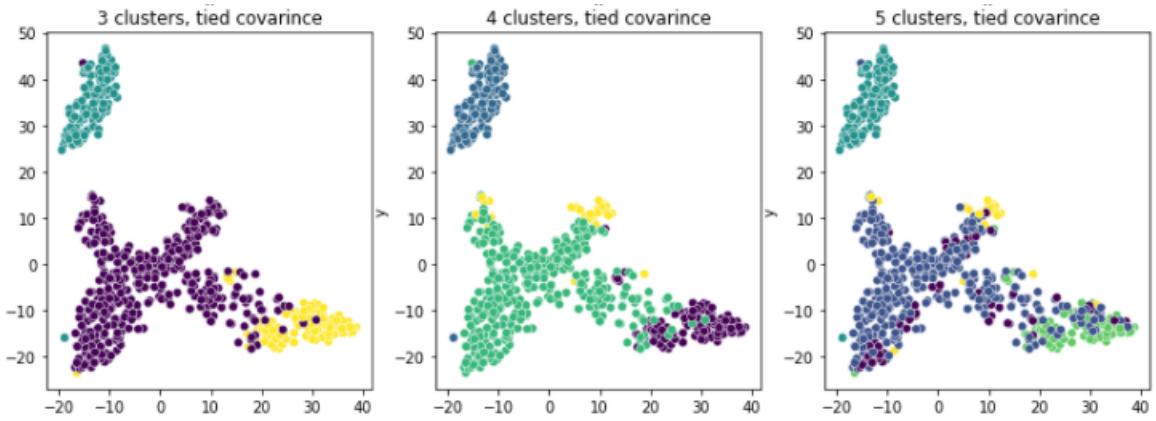
Rysunek 6: Skumulowana wyjaśniona wariancja

Okazało się, że już 3 komponenty wyjaśniają 85% wariancji, a 45 komponentów wyjaśnia 90%. Zdecydowaliśmy się więc użyć 45 komponentów. Następnie za pomocą T-SNE i podanych labeli tekstów zwizualizowaliśmy problem.



Rysunek 7: Wizualizacja właściwej klasteryzacji

Dobrze wyodrębniony jest klaster zawierający testy ze Starego Testamentu, Religie pochodzące z Azji są bardzo słabo odseparowane. Zastosowaliśmy algorytmy klastrowania, najlepszy okazała się GMM dla 3,4 lub 5 klastrów z covariancją 'tied'.



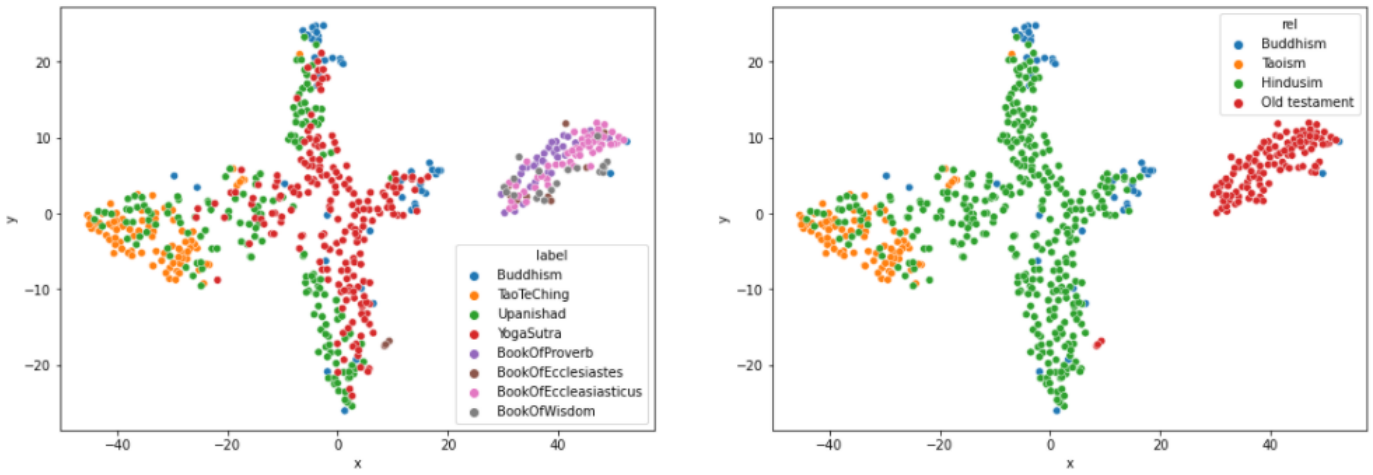
Rysunek 8: Wizualizacja klasteryzacji

clusters	covariance	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
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

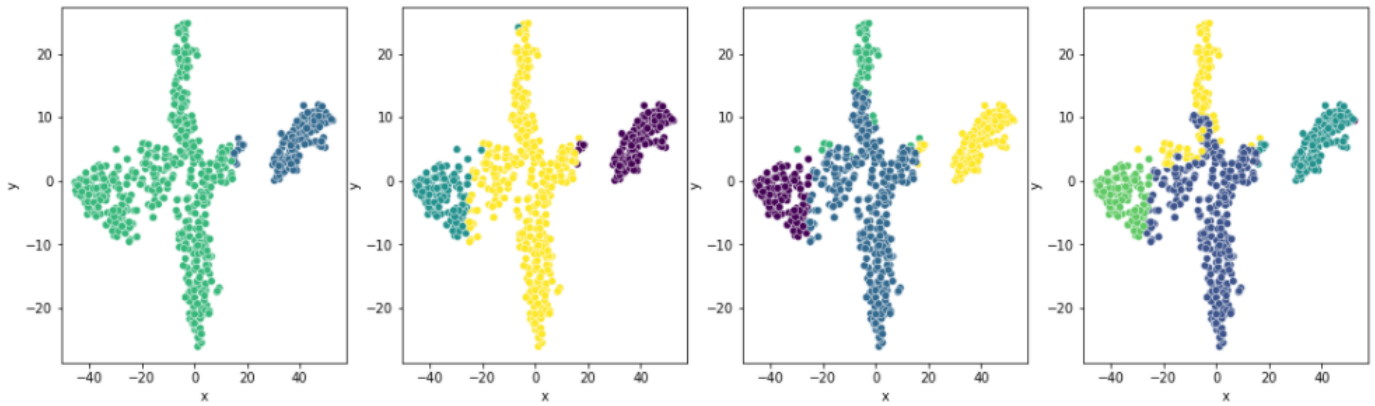
Rysunek 9: Metryki

3.4 SparsePCA

Spróbowaliśmy również SparsePCA - odmiany PCA dla macierzy rzadkich. Wyniki były jednak nieco gorsze, najepiej zadziałało KMeans.



Rysunek 10: Wizualizacja właściwej klasteryzacji



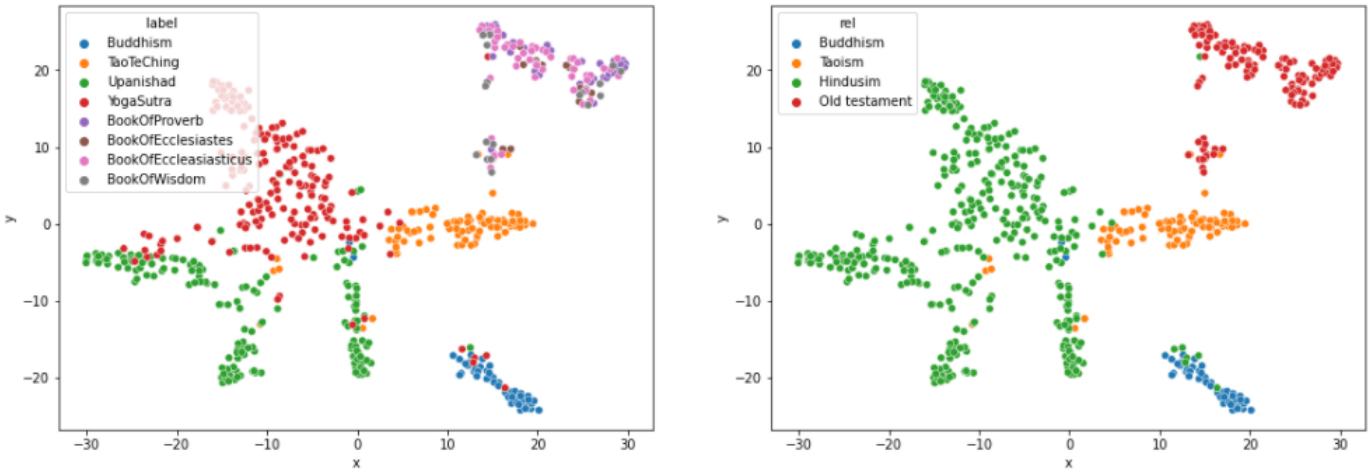
Rysunek 11: Wizualizacja klasteryzacji

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

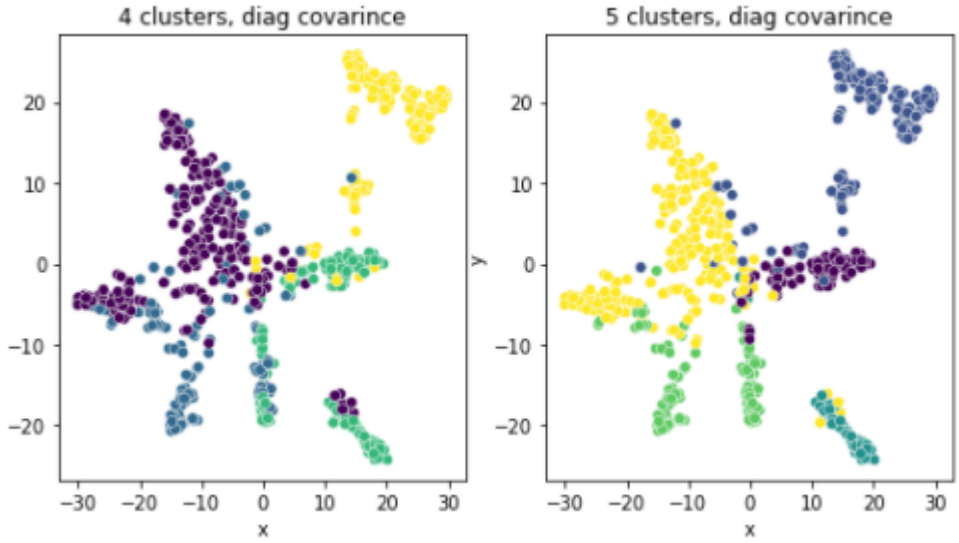
Rysunek 12: Metryki

3.5 NMF

Kolejną techniką, którą zastosowaliśmy do redukcji wymiarów było NMF - Non-negative matrix factorization, interpretowalny model działający na rzadkich macierzach o nieujemnych wartościach. Użyliśmy 8 komponentów i trenowaliśmy model na rzadkiej macierzy - csr_matrix. Wyraźnie oddzielone zostały klastry zawierające Księgi Starego Testamentu i Buddyzmu, dodatkowo całkiem wyraźnie oddzielne są Księgi Taoizmu. Niestety klasteryzacja osiągnęła słabe wyniki.



Rysunek 13: Wizualizacja właściwej klasteryzacji



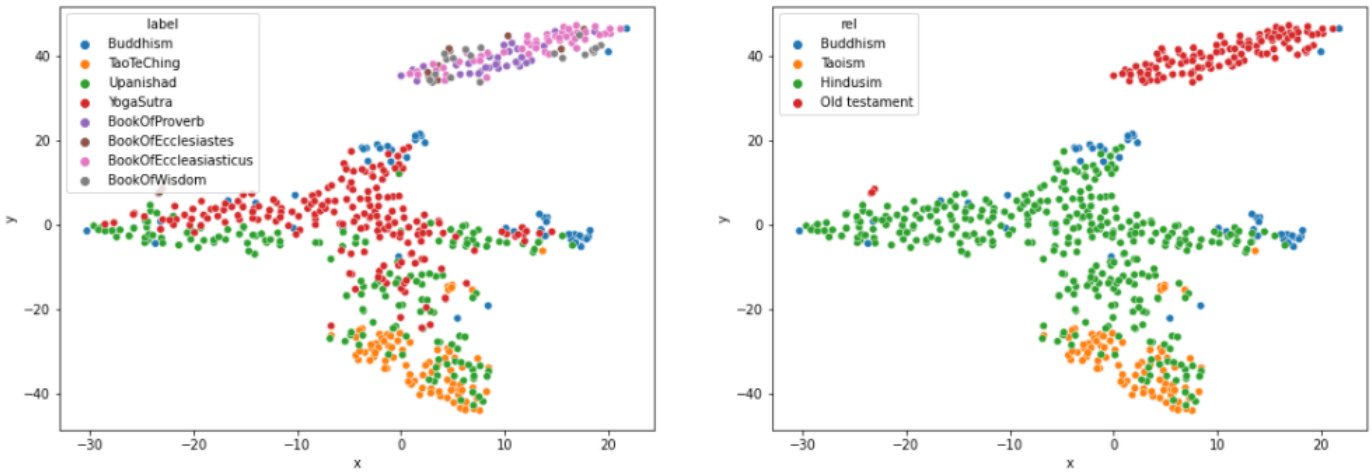
Rysunek 14: Wizualizacja klasteryzacji

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

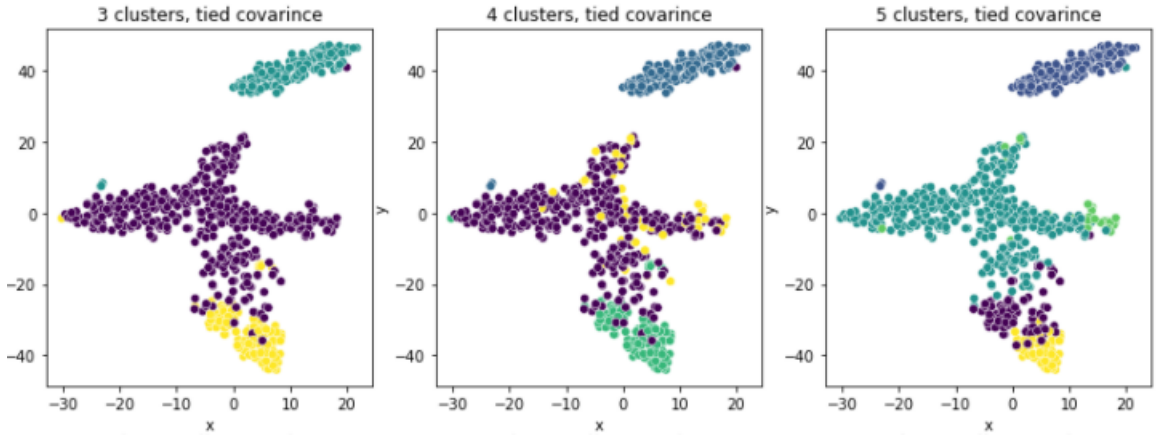
Rysunek 15: Metryki

3.6 TruncatedSVD

Ostatnią techniką redukcji wymiarów był algorytm TruncatedSVD, który również działa dla macierzy rzadkich. Wizualizacja przypomina tę wykonaną za pomocą PCA. Wyniki również ma bardzo podobn.



Rysunek 16: Wizualizacja właściwej klasteryzacji



Rysunek 17: Wizualizacja klasteryzacji

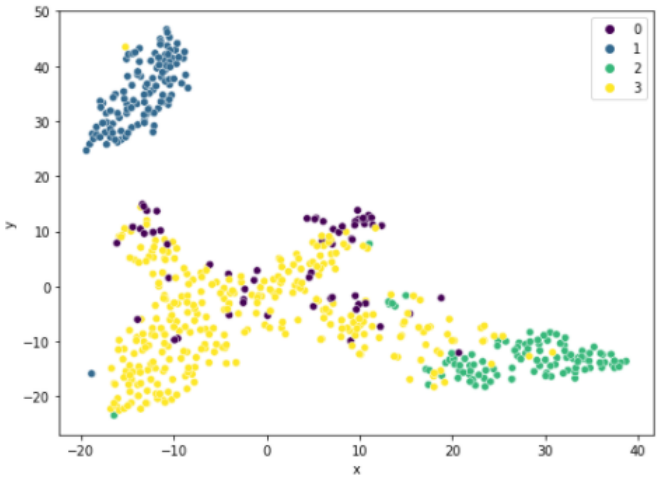
clusters	covariance	silhouette_score	davies_bouldin_score	rand_score	adjusted_mutual_info_score	mutual_info_score
3	tied	0.432942	0.825195	0.860551	0.741807	0.745419
4	tied	0.281062	1.504503	0.840902	0.680186	0.768905
5	tied	0.349039	1.082496	0.833559	0.626290	0.746307

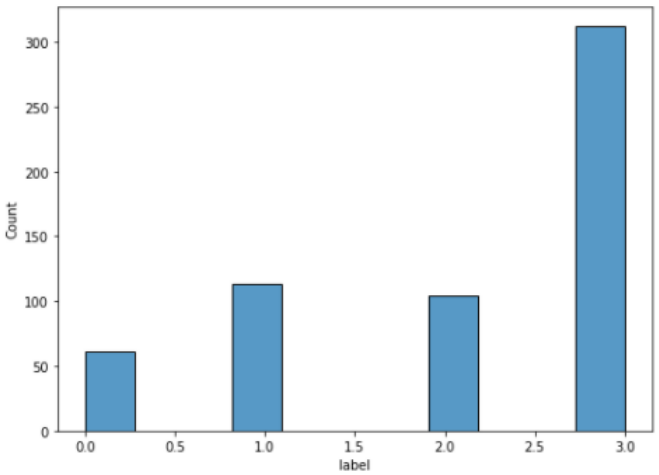
Rysunek 18: Metryki

4 Wyniki

Ostatecznie dwa najlepsze modele to

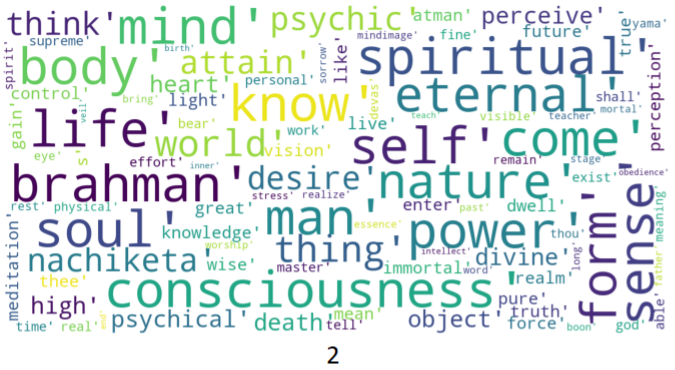
- GMM dla 4 klastrów z covariancją 'tied', na zbiorze zredukowanym przy pomocy PCA z 45 komponentami. W tym klastrowaniu widzimy, że jeden klaster jest zdecydowanie większy niż pozostałe, które są podobnej wielkości. W klastrze '0' rzucają się słowa o znaczeniu metafizycznym, np. spiritual, consciousness, soul, eternal itp. Natomiast w ostatnim przeważają zwykłe, codzienne słowa typu mind, great, state, appear.



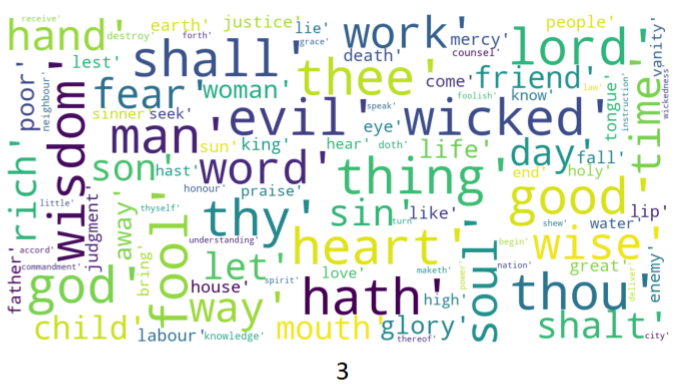


silhouette_score	0.412935
davies_bouldin_score	0.948276
rand_score	0.895347
adjusted_mutual_info_score	0.767429
mutual_info_score	0.838767
calinski_harabasz_score	364.309993

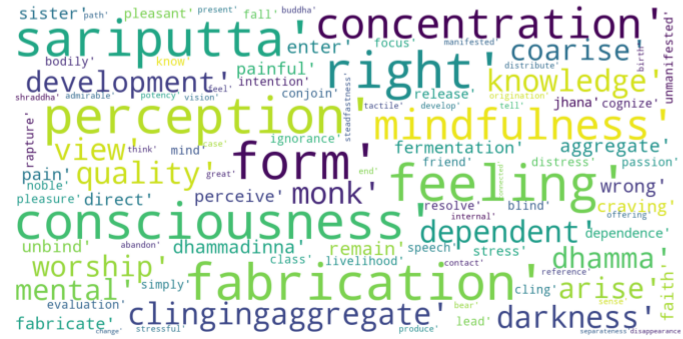
0



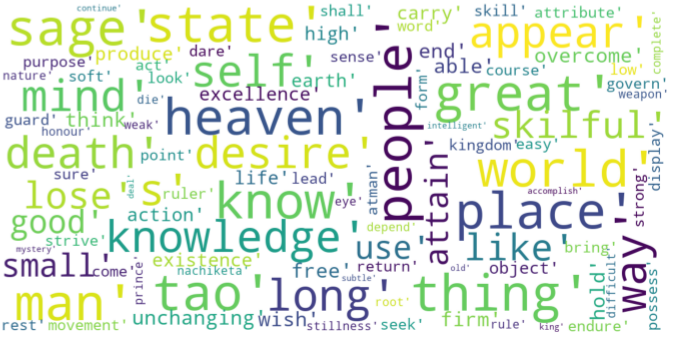
1



2

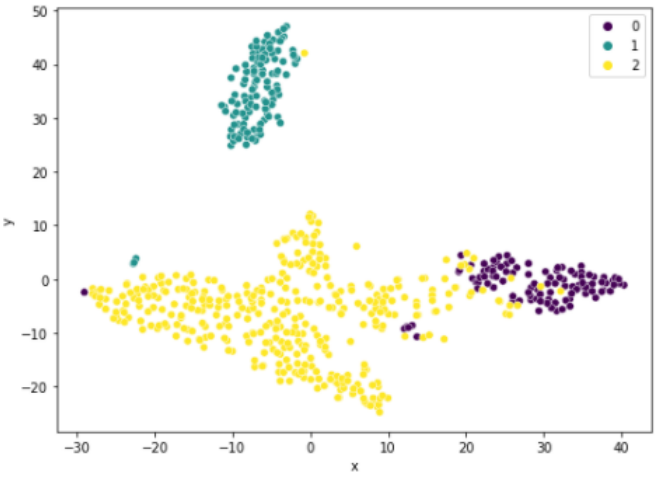


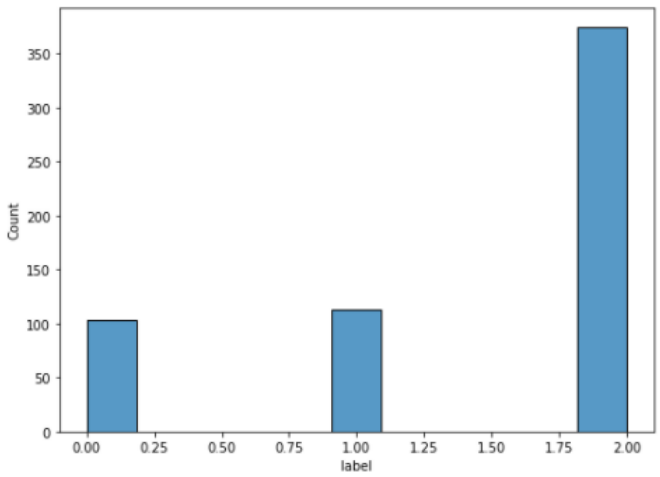
3



- GMM dla 3 klastrów z covariancją 'tied', na zbiorze zredukowanym przy pomocy TruncatedSVD z 50 komponentami. W tym przypadku również jeden klaster jest znacznie większy niż dwa pozostałe

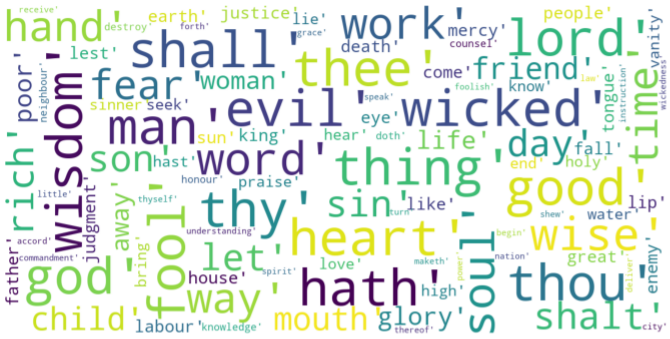
(niemalże równe). Patrząc na chmurę słów, uwagę przykuwa duża ilość metafizycznych określeń w klastrze ’2’, przyziemnych w klastrze ’1’ oraz brak jednolitości tematycznej w klastrze ’0’.





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



1

2