Wstęp do uczenia maszynowego: Projekt II
A study of Asian Religious and Biblical Texts

Paweł Koźmiński, Paulina Przybyłek, Ada Gąssowska

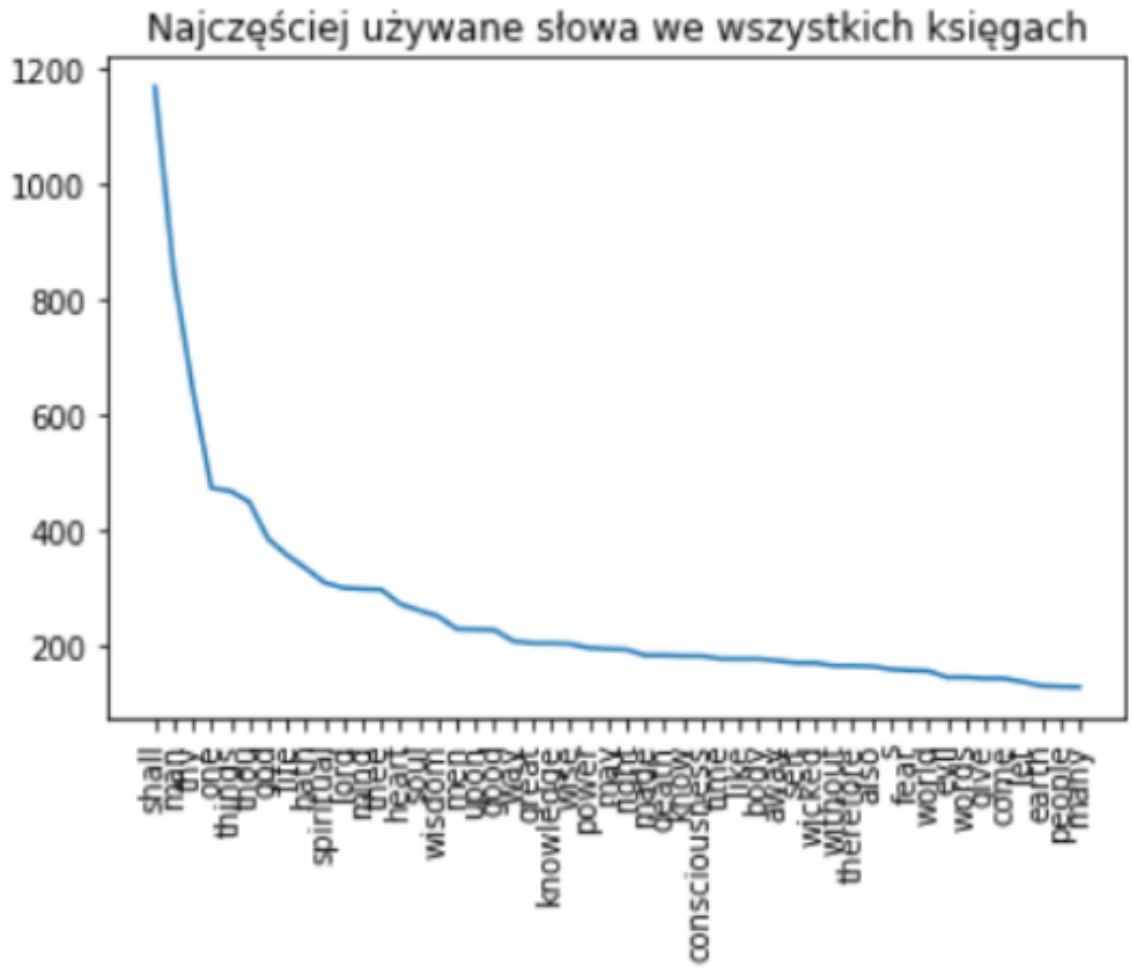
09.06.2020

Opis zbioru danych

- Ramka danych "All Books" zawiera informacje o słowach występujących w rozdziałach z ośmiu ksiąg religijnych z Azji.
- Ma ona nietypowe wymiary : 590 wierszy i 8267 kolumn.
- Każdy wiersz reprezentuje jeden rozdział jednej z ksiąg.
- Każda kolumna reprezentuje występowanie jednego słowa w rozdziałach ksiąg.
- W ramce wszystkie wartości to liczby całkowite większe lub równe zero, nie występują żadne braki danych.

Najczęściej wystepujące słowa

shall	1168
man	846
thy	645
one	473
things	468
thou	449
god	385
life	357
hath	334
spiritual	309



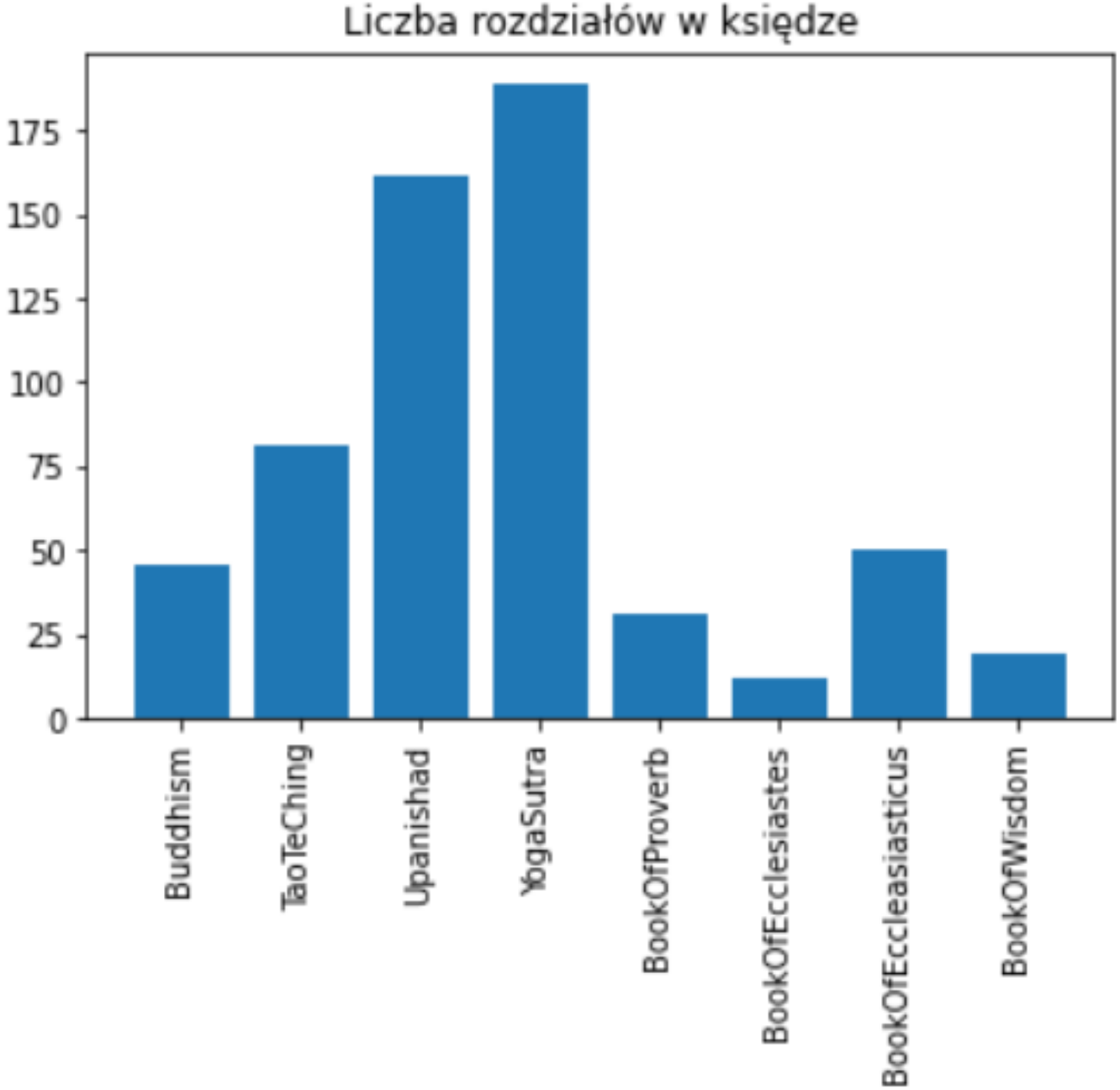
Najczęstsze słowo dla każdej księgi

label	Najczęściej używany wyraz:
BookOfEccleasiasticus	shall
BookOfEcclesiastes	shall
BookOfProverb	shall
BookOfWisdom	shall
Buddhism	right
TaoTeChing	tao
Upanishad	one
YogaSutra	spiritual

Występowanie słów

- Ponad połowa ze wszystkich słów występuje tylko w jednej z ośmiu ksiąg.
- Dla 60% słów ponad 90% ich wystąpień znajduje się w jednej z ksiąg.
- Tylko 1.3% ze wszystkich słów wystąpiło w każdej z ośmiu ksiąg (ich liczba wynosi 107).

Liczba rozdziałów w każdej z ksiąg



Unikalne słowa

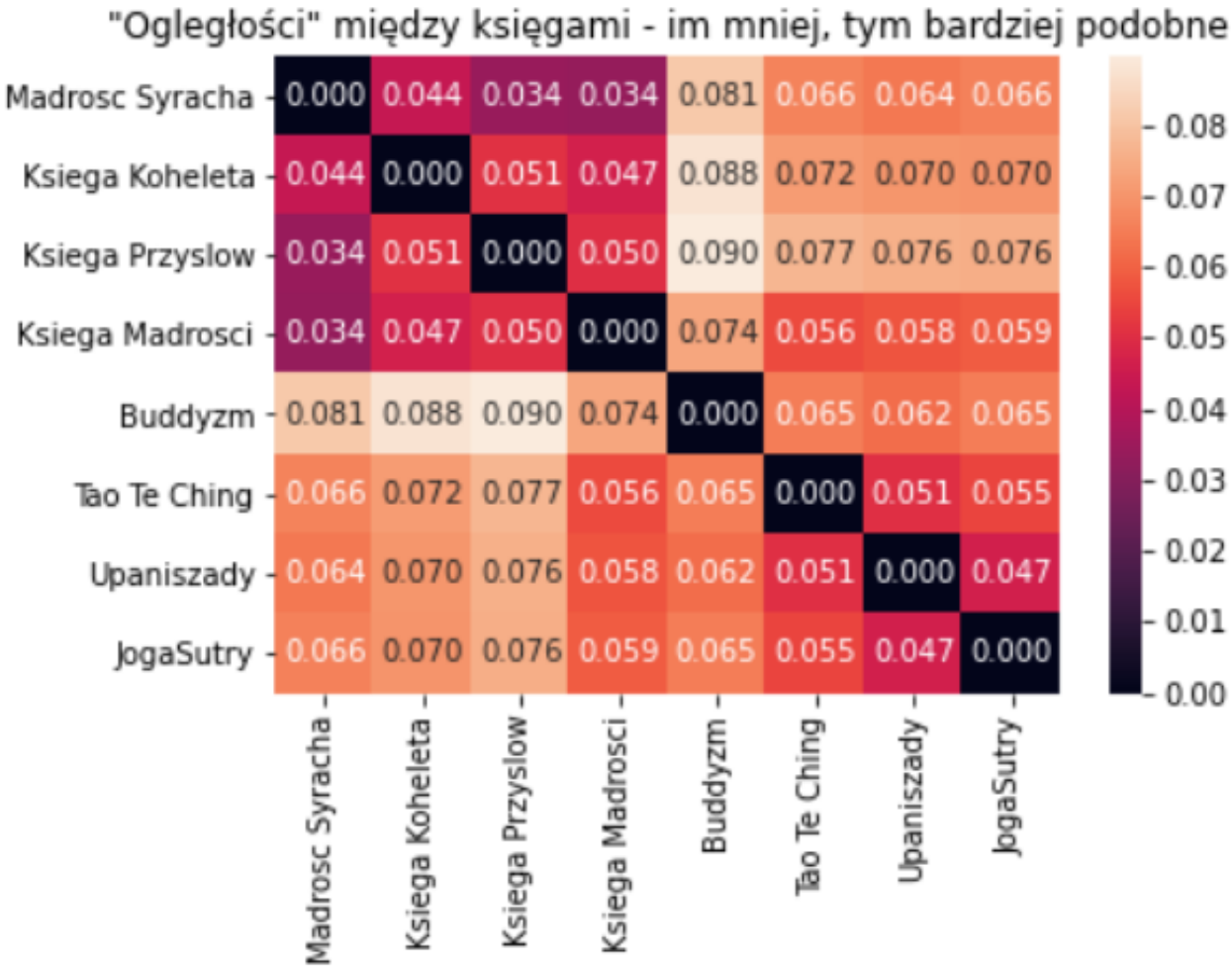
Liczba różnych słów w każdej z ksiąg:

BookOfEccleasiasticus	2995
BookOfEcclesiastes	970
BookOfProverb	1956
BookOfWisdom	1801
Buddhism	1348
TaoTeChing	1809
Upanishad	1849
YogaSutra	3282

Unikatowe słowa

label	Procent unikatowych słów w księdze
Madrosc Syracha	9.06%
Ksiega Koheleta	5.52%
Ksiega Przyslow	7.83%
Ksiega Madrosci	9.45%
Buddyzm	25.23%
Tao Te Ching	18.34%
Upaniszady	15.7%
JogaSutry	19.84%

Próba znalezienia odległości między księgami



Najbardziej podobnymi księgami okazują się Księga Przysłów i Mądrość Syracha oraz Księga Mądrości i Mądrość Syracha. Podobne są też do siebie księgi biblijne po prostu.

Próba lematyzacji

- Lematyzacja to sprowadzanie danego słowa do jego formy podstawowej
- Istnieją narzędzia do automatyzacji tego procesu, jednak jako że w naszej ramce jest sporo skomplikowanych słów, nie jesteśmy pewni czy narzędzie to poradziło sobie ze wszystkimi słowami.
- Po wykorzystaniu narzędzia do lematyzacji na naszym zbiorze, z 8266 słów dostaliśmy 6032.
- Przez liczność kolumn ciężko stwierdzić czy narzędzie zmieniło wszystkie słowa.

Próba analizy sentymentu

- Proste podejście - zliczenie słów o różnych sentymentach.
- Większość słów ma wydźwięk neutralny.
- Stosunek słów pozytywnych do negatywnych w każdej z ksiąg:

label	Negatywne/Pozytywnych
Madrosc Syracha	0.451429
Ksiega Koheleta	0.449198
Ksiega Przyslow	0.562771
Ksiega Madrosci	0.553571
Buddyzm	0.549296
Tao Te Ching	0.530556
Upaniszady	0.401766
JogaSutry	0.330022

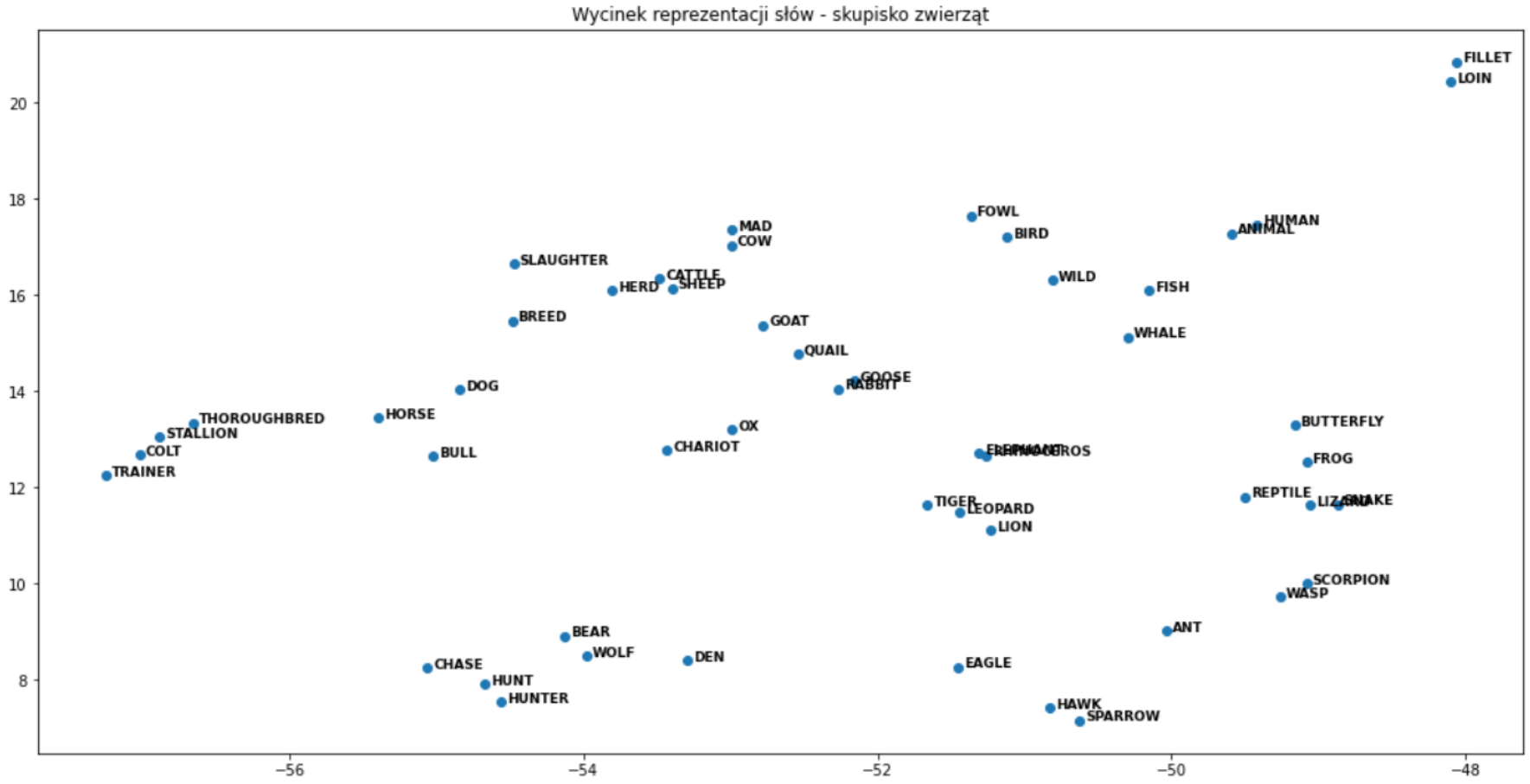
Inżynieria cech

Pierwszym etapem było przeprowadzenie lematyzacji, następnie ramkę przerobiliśmy przy pomocy trzech różnych narzędzi:

- Pierwszym z nich jest TfidfTransformer (term-frequency times inverse document-frequency).
- Jak sama nazwa wskazuje bazuje on na określaniu częstości występowania różnych słów.
- Po wykorzystaniu tego narzędzia otrzymujemy macierz rzadką, na której będziemy mogli testować nasze algorytmy.

Narzędzie Glove

- Zamiana na wektory, słowa o podobnym znaczeniu blisko siebie.



Ramki stworzone za pomocą Glove

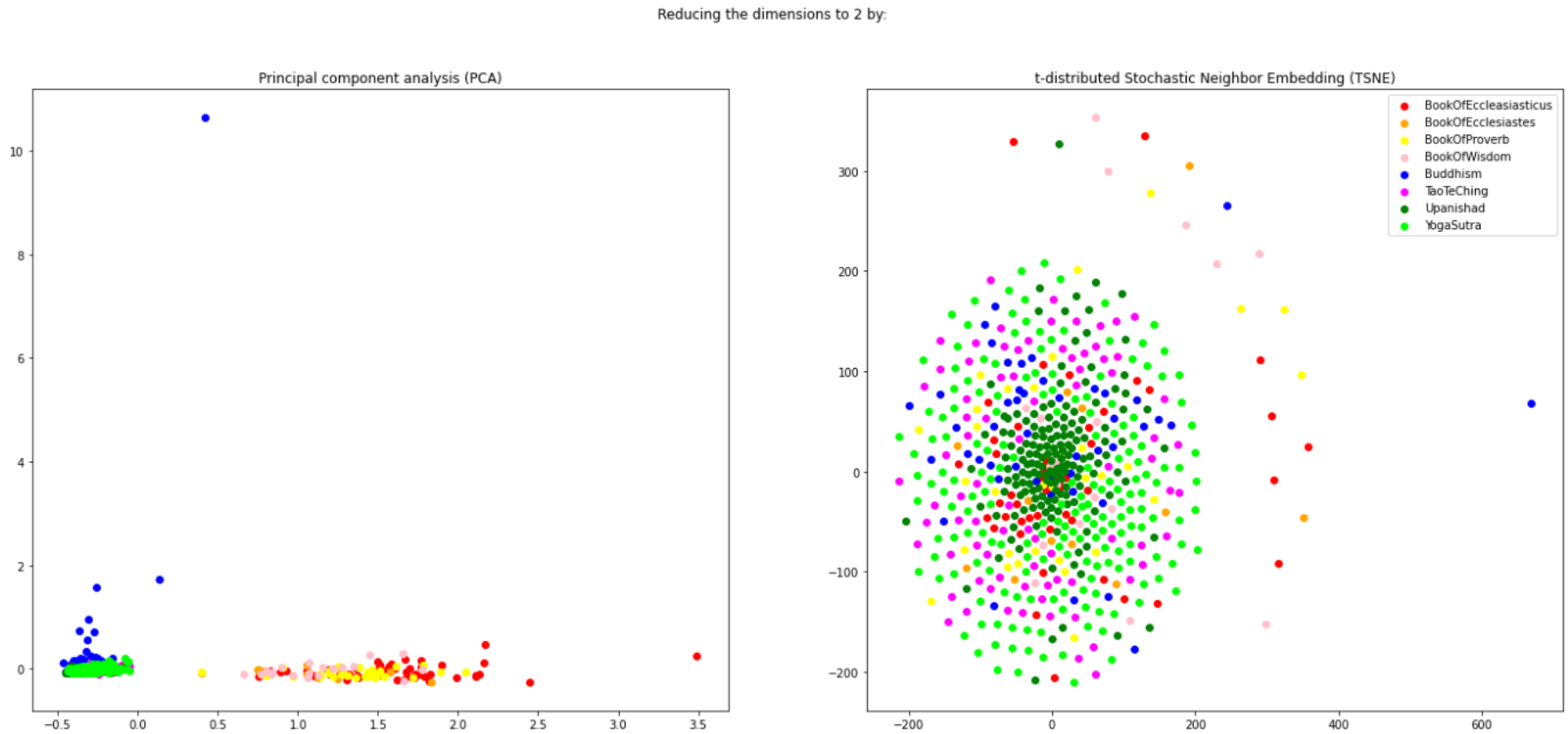
- Problemem, na który się natkneliśmy był fakt, że nie wszystkie słowa są znane przez Glove. Niestety jako że było ich 900 nie mogliśmy ręcznie tego poprawić, dlatego zostało tylko ich usunięcie.
- Pierwsza stworzonna ramka przy pomocy Glove, polega na stworzeniu stuelementowych wektorów dla każdego rozdziału (wiersza).
- Drugi wykorzystany sposób stworzenia ramki z pomocą Glove to wybranie kilku najbardziej popularnych słów w rozdziale i skonkatenowanie wektorów ich reprezentujących.

Wybór liczby klastrów

- Stworzyliśmy trzy ramki: df_tfidf, df_glove_remove_1 i df_glove_remove_2.
- Dla każdej z nich szukaliśmy liczby klastrów za pomocą trzech miar: **Silhouette, Davies-Bouldin i Calinski-Harabasz**.
- W większości przypadków najlepsze wyniki otrzymaliśmy dla podziału na dwa klastry. Trzy/cztery też miały dość dobre wyniki.

Redukcja wymiarów

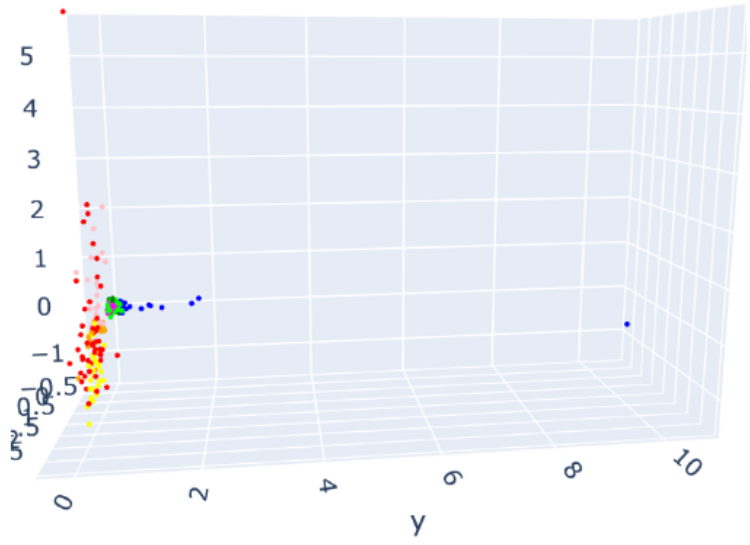
- Oryginalne dane przeskalowaliśmy za pomocą MinMaxScalera i logarytmicznej normalizacji.
- Do redukcji wymiarów użyliśmy PCA oraz TSNE
- Dla dwóch i trzech wymiarów spróbowaliśmy przedstawić nasze dane (przeskalowane oryginalne oraz te przerobione podczas inżynierii cech) na rysunkach:



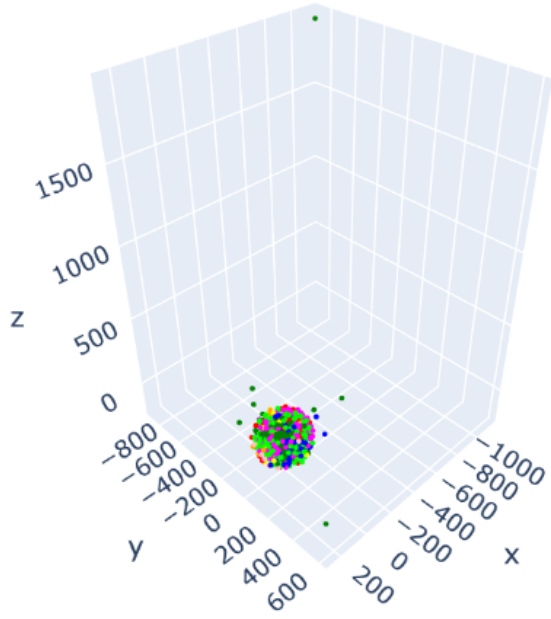
Redukcja do trzech wymiarów

Original data: Reducing the dimensions to 3 by:

Principal component analysis (PCA)

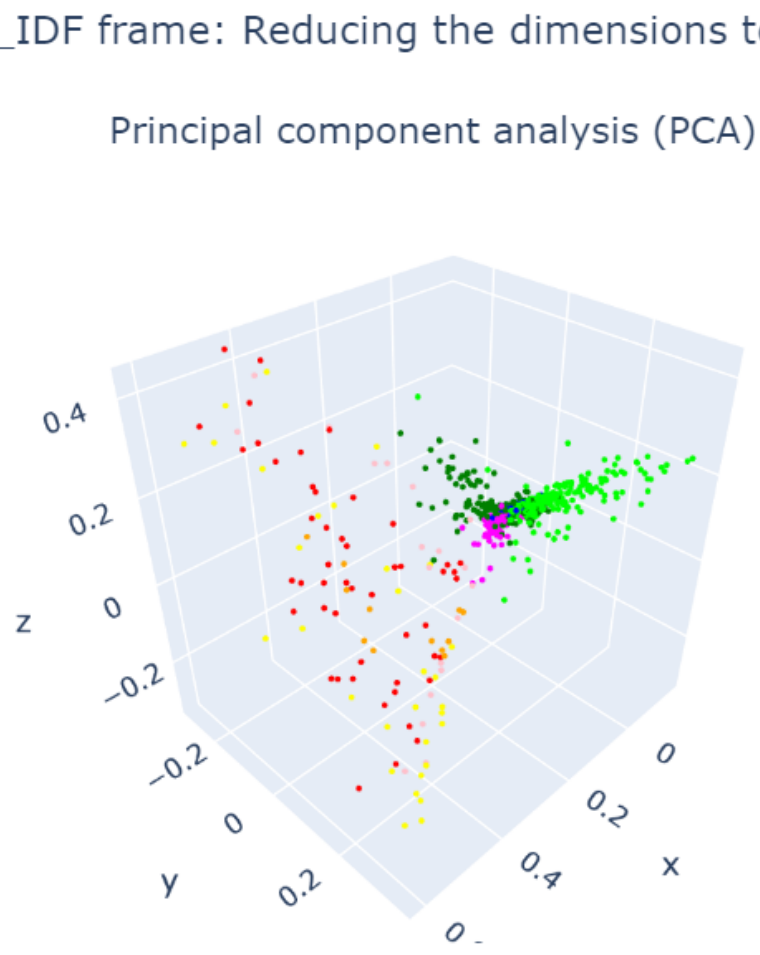


t-distributed Stochastic Neighbor Embedding (TSNE)



Najlepszy obiektywnie podział dla 3 wymiarów

Dla ramki stworzonej przy pomocy TFIDF, redukując wymiary do trzech za pomocą PCA otrzymaliśmy rysunek, na którym najbardziej widoczny jest podział.



Wypróbowanie algorytmu DBSCAN

- Sprawdziliśmy działanie algorytmu DBSCAN, na naszych ramkach.
- Parametr min_samples ustawiliśmy jako wymiar danych zwiększony dwukrotnie.
- Parametr eps ustawiliśmy jako odległość, względem której większość (tzn. 95%) spośród k-tych sąsiadów dla poszczególnych obserwacji jest bliżej.
- Dla większości wymiarów algorytm zwrócił tylko jeden klaster. Tylko dla dwóch wymiarów,w niektórych przypadkach, algorytm podzielił zbiór na sześć klastrów.
- Wyższe wyniki Silhouette (ok. 0.5) zostały osiągnięte dla redukcji PCA. Dla wszystkich ramek były one podobne.

Inne algorytmy na ramce bez redukcji wymiarów - podsumowanie

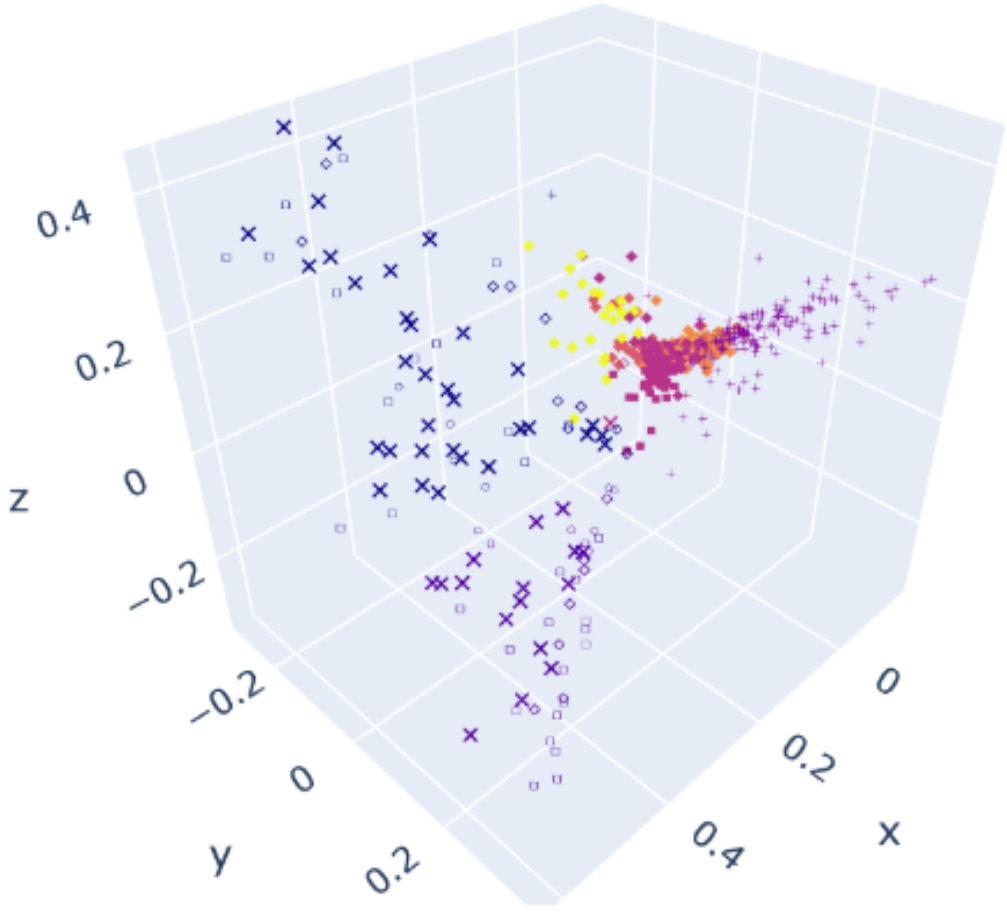
- Na ramkach bez redukcji wymiarów wypróbowaliśmy 6 algorytmów:
  - K Means
  - Mini Batch K Means
  - Birch
  - Agglomerative Clustering - różne połączenia (Warda, kompletne, pojedyncze)
- Dla każdej ramki, dla każdej metody sprawdziliśmy ile wynosi miara silhouette w zależności od liczby klastrów (od 2 do 8).
- Wyniki okazały się być bardzo słabe. Najwyższy wynik (ok. 0.5) osiągnęło Agglomerative Single na ramce df_glove_remove_1 dla dwóch klastrów.

Algorytmy na ramkach z redukcją wymiarów

- Przy pomocy miar silhouette oraz adjusted_mutual_info i adjusted_rand (korzystających z przypisanych etykiet) zbadaliśmy działanie sześciu wspomnianych wczesniej algorytmów klasteryzujących dla różnych liczby zredukowanych wymiarów (redukcja zdecydowanie podniosła efektywność algorytmów) oraz klastrów.
- Po redukcji TSNE klasteryzacja na pierwszej ramce glove osiągała wysokie wyniki silhouette - skupienia zdecydowanie różniły się od siebie, lecz wyniki miar korzystających z prawdziwych etykiet nie były najlepsze
- Po podziale etykiet na dwie grupy - księgi bliskiego i dalekiego wschodu, uzyskaliśmy bardzo wysoką skuteczność przy klasteryzacji ksiąg. Agglomerative Clustering z metodą Warda na ramce tfidf niepoprawnie przypisał tylko jedną obserwację (adjusted_rand prawie równe 1)
- Podział na 8 ksiąg nie przyniósł tak dobrego rezultatu.

Podział na 8 klastrów

Z podziałem na 8 skupień najlepiej poradził sobie algorytm aglomeratywny z metodą Warda po PCA do 8 wymiarów na ramce tfidf.



Podział na 2 klastry – najlepsze osiągnięte wyniki

Z podziałem na dwa skupienia (wyższe wyniki Silhouette niż dla ośmiu) najlepiej poradził sobie również algorytm aglomeratywny z metodą Warda po PCA do 11 wymiarów na ramce tfidf.