# Projekt dot. klasteryzacji - raport końcowy

Paulina Przybyłek
Ada Gąssowska
Paweł Koźmiński

21 kwiecień 2020

**Abstrakt.** Projekt dotyczy klasteryzacji na zbiorze o księgach i tekstach religijnych i wybraniu najlepszego podziału (za najlepszy ,oglibyśmy uznać 8, gdyż tyle danych o różnych księgach zabiera zbiór). Skupia się on na trzech obszarach: analizie zbioru i danych o księgach, inżynierii cech i przekształcenie zbioru jak najlepiej dla modelowania oraz modelowanie i wybór najlepszego podziału na klastry na podstawie kilku różnych miar i metod. Dodatkowo w projekcie duży nacisk będzie kładziony na zmniejszenie zbioru - zmianę formy (m.in. wektory słów) czy redukcję wymiarów.

# 1 Wstęp

Cała analiza w projekcie oparta jest na zbiorze **A study of Asian Religious and Biblical Texts**, który można znaleźć i pobrać ze strony **link**. Zbiór danych zawiera 8265 słów występujących w ośmiu księgach religijnych przygotowanych jako mini-korpus tych ksiąg (gdyż nie są to wszystkie słowa, a wybrane przez twórców). Większość świętych tekstów w tym zbiorze danych zebrano z projektu Gutenberg. Dane są ułożone w następujący sposób: kolumny to słowa a wiersze to rozdziały, a w każdej komórce mamy liczbę wystąpień danego słowa w tym rozdziale. Przez zbiór ma nietypowe rozmiary - 590 wierszy i 8266 kolumn. W wykorzystywanej przez nas ramce wszystkie wartości to liczby całkowite większe lub równe zero, nie występują żadne braki danych.
Na podstawie tego zbioru wykonano analizę zbioru (wykorzystując etykiety określające co to za rozdział i księga), natomiast do klasteryzacji, które należy do uczenie nienadzorowane, usunięto etykiety nazw ksiąg. Oczywiście zbiór został wcześniej do tego specjalnie przygotowany, a modelowanie opierało się na sprawdzeniu kilku możliwości inżynierii cech danych, aby określić najlepszy do tego sposób.

# 2 EDA - wstępna analiza zbioru

EDA polega głównie na zapoznaniu się z danymi i wyciągnięciu wniosków, które mogą pomóc przy późniejszych etapach. Przypominając, zbiór zawiera 8265 słów występujących w ośmiu księgach religijnych, które zostały poniżej krótko opisane.

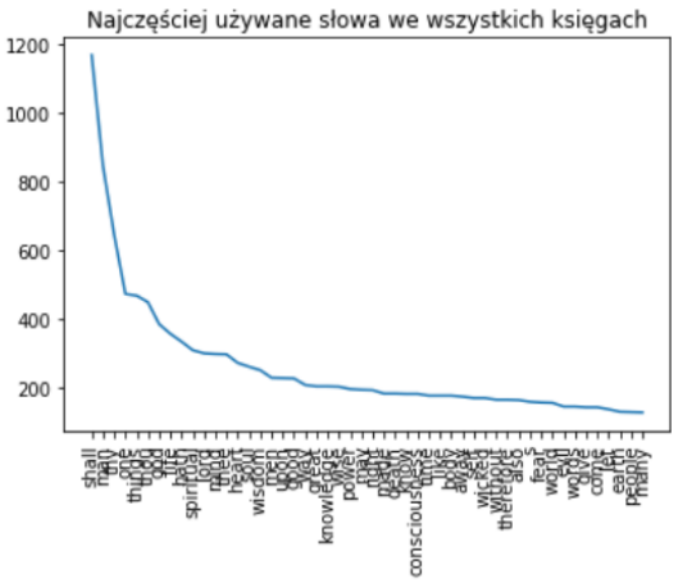
- **Book of Eccleasiasticus** - Mądrość Syracha - jedna z ksiąg deuterookanonicznych Starego Testamentu (czyli takich, które są w Biblii chrześcijańskiej, ale nie w hebrajskiej. Napisana ok. 190 r.p.n.e. w Jerozolimie.
- **Book Of Ecclesiastes** - Księga Koheleta - znajduje sie w obu Bibliach. Datowanie księgi nie jest pewne, choć są znaki, aby określać je na III w.p.n.e.
- **Book of Proverb** - Księga Przysłów - również w obu Bibliach, jest pracą zbiorową złożoną z różnych tekstów przez nieznanego autora ok. V w.p.n.e.
- **Book of Wisdom** - Księga Mądrości - napisana prawdopodobnie w Aleksandrii (Egipt) - datowana na ok. 50 r.p.n.e.
- **Buddhism** - oczywiście nazwa religii dalekiego wschodu - buddyzmu, który najczęściej jest wyznawany w kraja półwyspu indochińskiego, Malezji, Chinach i Monoglii.
- **Tao Te Ching** - także: Lao Tsu - chińska księga najprawdopodobniej napisana w 6 wieku p.n.e. przez mędrca Laozi. Uważana jest za podstawowe dzieło taoizmu - jedną z najpopularniejszych chińskich religii. Badacze twierdzą także, że miała ona wpływ także na kształtowanie filozofii buddystów.

- **Upanishads** - upaniszady - teksty (w liczbie ponad 200) religii hinduizmu należące do wedyjskiego objawienia o treści religijno - filozoficznej. Jedne z późniejszych tekstów należących do Wed (świętych ksiąg hinduizmu), napisane w VIII - III w.p.n.e. Są do dziś wykorzystywane nawet w naszej kulturze wśród osób praktykujących medytację.

- **Yoga Sutra** - jogasutry (hinduizm) - najstarszy tekst klasycznej Jogi (w znaczeniu systemu filozofii indyjskiej, uznającego autorytet Wed), składający się ze 195 sutr - zwięzłych sentencji. Według hinduskich tradycji, ich autorem był żyjący w II w.p.n.e. Patañjāli.

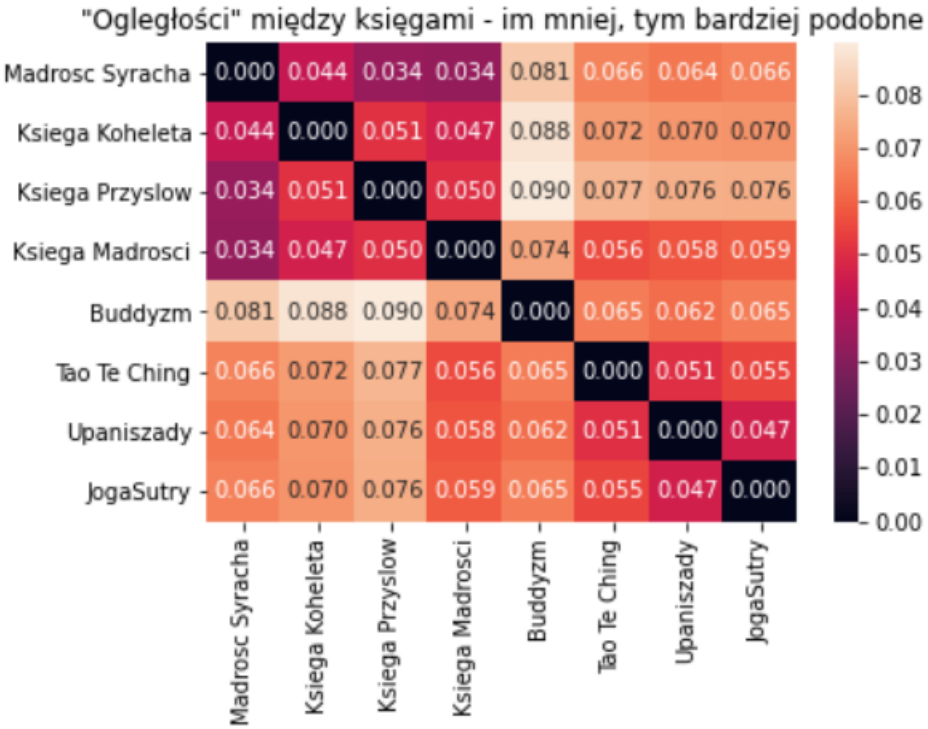
Mając liczbę wystąpień słów dla każdego rozdziału w danej księdze, przeprowadziliśmy dwie analizy - dla rozdziałów i ogółem dla ksiąg. W zbiorze nie ma tzw. *stopwords*, więc nie trzeba było go czyścić i od razu mogliśmy przejść do wyciągania ciekawych wniosków. Sprawdziliśmy: najczęściej występujące słowa, średnią liczbę użyć danego słowa, liczbę słów, występowanie słów, w tym unikatowość czy najczęstsze części mowy, analizę sentymentu oraz podobieństwo ksiąg. Każde działanie zwróciło ciekawy wynik, jednak przytoczymy najważniejsze wnioski z tego etapu. Na poniższym wykresie oraz w tabeli mamy przedstawione najczęściej występujące słowa ogółem.

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




Jak widzimy, dwa-trzy słowa się wyróżniają a potem mamy wyraźny spadek. Większość słów nie przekracza nawet 200 wystąpień. Najczęściej występującym słowem jest **shall**. Powtórzyło się ono również jako najpopularniejsze w czterech księgach, natomiast słowa z księgi Buddyjskiej i TaoTeChing nie znalazły się w 10 najpopularniejszych słowach dla ogółu, więc możlwie, że przejawiają się one w większośći właśnie tam (tao szczególnie). Okazuje się, że tao występuje tylko w księdze TaoTeChing, a słów unikatowych, których 90% występień jest tylko w jednej księdze to aż ponad 60% wszystkich słów, które mamy w korpusie. Z drugiej strony tylko 1.3% (107) słów występuje w każdej z ksiąg i są to przeważnie rzeczowniki. Wiadomo, że każda księga ma

inną długość i liczbę słów czy rozdziałów, dlatego sprawdziliśmy procent unikatowych słów w księdze i najwięcej, bo aż 25.23% słów, które występują w księdze Buddyzmu jest unikatowym. Najdłuższą księgą okazała się Mądrość Syracha, a o największej liczbie rozdziałów Jogasutra. Co ciekawe jedna księga Buddyzmu nie ma żadnego ze słów, które występują w korpusie. Postanowiliśmy też sprawdzić, czy księgi religijne są do siebie podobne i zrobiliśmy to, biorąc pod uwagę podobieństwa występień słów do siebie. Poniżej przedstawiono heatmapę rezultatów.




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. Oczywiście musimy pamiętać, że korzystamy tylko z części słów występujących naprawdę w księgach. EDA postanowiliśmy zakończyć analizą sentymentu i lematyzacją, czyli zmianą słowa do jego podstawowej formy. Okazuje się, że wszystkie księgi są neutralne, jednak bardziej pozytywne niż negatywne i to przeważnie o jakieś 2 razy. Natomiast po zastosowaniu lematyzacji liczba słów zmniejszyła się o ponad 2 tys. słów.

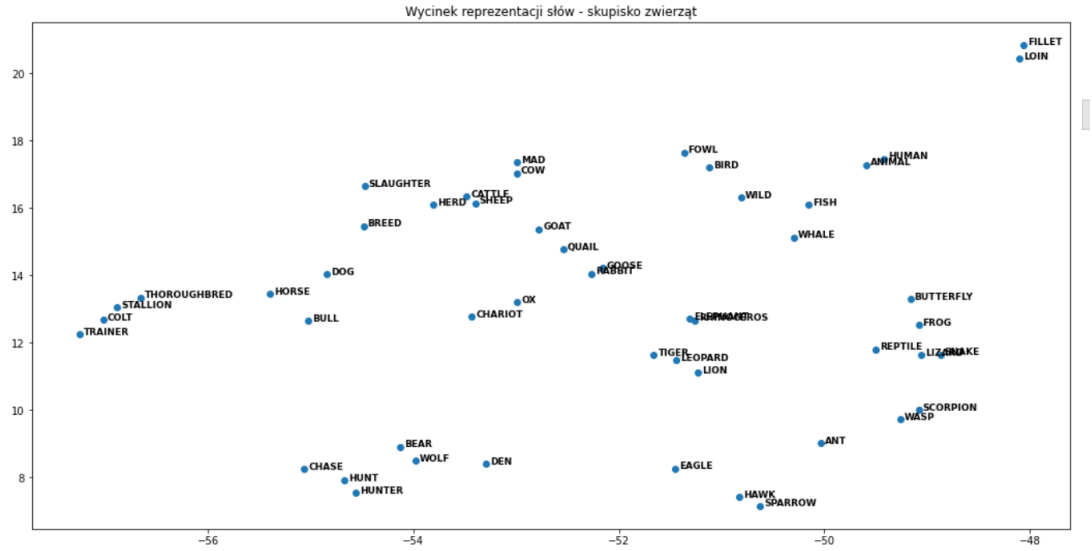
# 3 Inżyniera cech

Krótko przedstawiając ten punkt: ramkę danych poddano lematyzacji i zmniejszono ją aż o 2234 słowa, przygotowanie ramek danych na dwa sposoby:

- skorzystanie z narzędzia TfidfTransformer (term-frequency timesinverse document-frequency). Jak sama nazwa wskazuje bazuje on na określaniu częstościwystępowania różnych słów.

- skorzystanie z embeddingu GloVe na dwa sposoby:
  - korzystając z możliwości dodawania wektorów reprezentujących słowa spróbujemy obliczyć średnią ważoną z wektorów dla każdego rozdziału, gdzie wagami będą znormalizowane liczby występowania poszczególnych słów.
  - poprzez wybranie kilku najbardziej popularnych słów w księdze (pomoże nam brak stopwords) i skonkatenowania wektorów ich reprezentujących.

W ten sposób powstały nam trzy ramki danych przeznaczone do kolejnego etapu - modelowania. Pierwsza to macierz rzadka powstała dzięki TfidfTransformer. Druga i trzecia to rezultaty wykorzystania embedingu. GloVe zamienia słowa na wektory, aby uwypuklić podobieństwo słów. Zobaczmy to na przykładzie.

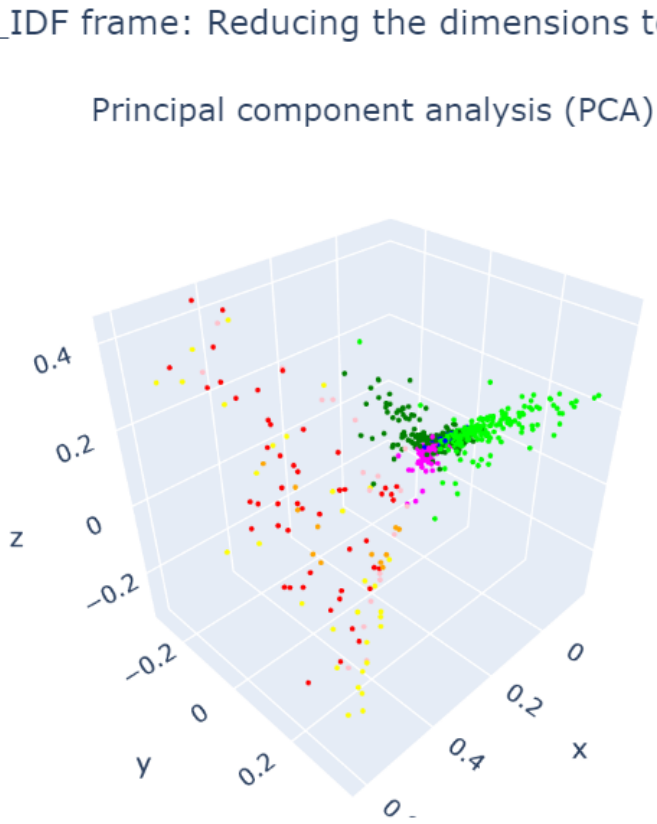



Możemy zauważyć jak podobne do siebie zwierzęta są w blisiej odległości od siebie - jak leopard, tygrys i lew. Jednak w słowniku GloVe nie ma wszystkich słów, które występują w naszych danych (m.in. archaizmów). Zdecydowaliśmy się usunąc te słowa, gdyż było to 3% sumy wszystich słów i aż 68% to słowa występujące w jednej z ksiąg. Pierwsza ramka zawiera stuelementowe wektory dla każdego rozdziału (wiersza), druga powstała w poprzez wybranie kilku najbardziej popularnych słów w rozdziale i skonkatenowanie wektorów ich reprezentujących. Po zmniejszeniu wymiaru ramek uznaliśmy, że można przejść do kolejnego etapu.

# 4 Modelowanie

Wstępne modelowanie opierało się na znalezieniu optymalnej liczby klastrów dla przygotowanych na trzy sposoby danych, wykorzystując domyślne parametry i metodę kombinatoryczą - K-średnich oraz metodę hierarchiczną - aglomeracyjną. Miary, które pomogły nam wybrać to: metoda łokcia, metoda Silhouette, indeks Daviesa-Bouldina, indeks Calińskiego-Harabasza. W większości przypadków najlepsze wyniki otrzymaliśmy dla podziału na dwa klastry. Chociaż trzy/cztery też miały czasem dość dobre

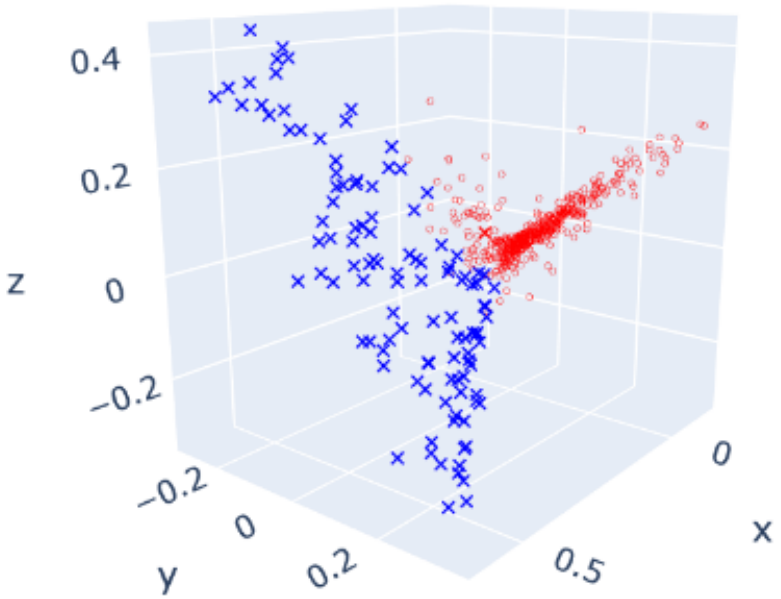
wyniki. Spodziewaliśmy się podziału na osiem, bo tyle różnych ksiąg mamy, jednak tak się nie stało. Pojawiło się pytanie czy to może podział na dwa skupiska to po prostu podział na religie Bliskiego i Dalekiego Wschodu? Zanim przeszliśmy do wyciągania wniosków chcieliśmy wykonać cały proces modelowania.
Postanowiliśmy wykonać inne podejście - redukcję wymiarów przy użyciu Principal component analysis (PCA) i t-distributed Stochastic Neighbor Embedding (tSNE). Użyliśmy tutaj trzech stworzonych uprzednio ramek danych oraz orginalnej ramki . Przed zastosowaniem metod redukujących istotna była normalizacja zmiennych na orginalnej ramce: MinMaxScaler + LogNormalizacja. Sprawdziliśmy wykresy redukcji wymiarów w dwóch oraz trzech wymiarach. Niestety, działanie tSNE w trzech wymiarach nie skutkowało rozdzieleniem klastrów dla którejkolwiek z ramek. Jednak subiektywnie można stwierdzić, że dla wszystkich wykonanych powyższych prób, najlepiej spisał się PCA na ramce po transformacji TfidfTransformer (wykres poniżej efekt tego).




Następnie postanowiliśmy sprawdzić działania algorytmu DBSCAN, ustawiając własne parametry. Dla większości wymiarów algorytm zwrócił tylko jeden klaster. Tylko dla dwóch wymiarów, w niektórych przypadkach, algorytm podzielił zbiór na sześć skupisk. Wyższe wyniki Silhouette (ok. 0.5) zostały osiągnięte dla redukcji PCA.
Poprzez niezadowalające wyniki sięgneliśmy po inne algorytmy, dla których można wyraźnie określić liczbę klastrów do podziału:

1. K Means
2. Mini Batch K Means
3. Birch
4. Agglomerative Clustering (wykorzystując różne połączenia: Warda, kompletne, pojedyncze)

Wykorzystaliśmy ramki bez redukcji wymiarów i mierzyliśmy je miarą silhouette. Wyniki okazały się być bardzo słabe. Najwyższy wynik (ok. 0.5) osiągnęło Agglomerative Single na ramce ramce GloVe, złożonej ze stuelementowych wektorów dla każdego wiersza, dla dwóch klastrów.
Ostatnim podejściem było przy pomocy miar silhouette oraz adjusted_mutual_info i adjusted_rand (korzystających z przypisanych etykiet) zbadaliśmy działanie sześciu wspomnianych wcześniej algorytmów klasteryzujących dla różnej liczby zredukowanych wymiarów (redukcja zdecydowanie podniosła efektywność algorytmów) oraz klastrów. Po podziale etykiet na dwie grupy - księgi Bliskiego i Dalekiego Wschodu, uzyskaliśmy bardzo wysoką skuteczność. Agglomerative Clustering z metodą Warda na ramce po transformacji TfidfTransformer niepoprawnie przypisał tylko jedną obserwację (adjusted_rand prawie równe 1). Wynik można zobaczyć na wykresie poniżej. Jednak podział na 8 skupisk - liczba ksiąg w danych nie okazał się dobrym podziałem, algorytmy nie podzieliły tak jakbyśmy tego chcieli.



# 5 Wyniki i konkluzje

Krótko podsumowując, zastosowaliśmy trzy sposoby przetworzenia danych - transformację TfidfTransformer oraz embeding przy użyciu Glove. Przy modelowaniu spróbowaliśmy także zredukować wymiar tych danych dzięki PCA oraz tSNE. PCA okazało się być lepszym narzędziem zarówno do wizualizacji jak i klasteryzacji. Najlepszy podział to dwa skupiska - religie Bliskiego i Dalekiego Wschodu. Agglomerative Clustering z metodą Warda na ramce po transformacji TfidfTransformer przypisał najlepiej rozdziały do tych skupisk.