Raport
Uczenie nienadzorowane
Projekt

Adrianna Grudzień, Mateusz Stączek

8 czerwca 2021 r.

Spis treści

1 Wprowadzenie

W tym projekcie pracowaliśmy nad ośmioma tekstami religijnymi:

- Upanishads,
- Yoga Sutras,
- Buddha Sutras,
- Tao Te Ching,
- Księga Mądrości,
- Księga Przysłów,
- Księga Koheleta I,
- Księga Koheleta II.

Każdy tekst został podzielony na rozdziały (łącznie 590 rozdziałów), a celem projektu było podzielenie tych rozdziałów na grupy odpowiadające księgom oryginalnego pochodzenia (mimo, że wszystkie teksty były na tematy religijne, to jak widać na Rysunku 1, są między nimi pewne różnice).

Dane pochodzą ze strony archive.ics.uci.edu.

Rysunek 1: Chmury słów z kilku przykładowych rozdziałów.

2 Inżynieria cech

Ponieważ oryginalnymi danymi był surowy tekst, postanowiliśmy na jego podstawie stworzyć własną ramkę danych zawierającą kluczowe informacje (głównie statystyczne) o poszczególnych rozdziałach.

Uwzględniliśmy następujące zmienne:

- długość rozdziału
- liczba zdań
- średnia długość zdania
- liczba słów (bez znaków przystankowych)
- liczba unikalnych słów (bez znaków przystankowych)

- średnia długość słów
- poziom skomplikowania rozdziału (wskaźnik FRE)
- zabarwienie rozdziału - pozytywne/negatywne/neutralne (polarność)
- subiektywność rozdziałów (tekst subiektywny - 1, obiektywny - 0)

Duża liczba uzyskanych wartości miała szeroki zakres, dlatego też przeprowadziliśmy ich normalizację. Dodatkowo z powodu obecności wartości odstających, skorzystaliśmy z Robust Scaler, a nie Standard Scaler. Ponadto do powyższych danych "dokleiliśmy"BoW (Bag of Words, wyjaśnione dalej) dla około 4000 słów.

3 Modelowanie i wizualizacja wyników

Eksperymenty przeprowadziliśmy na następujących zbiorach danych:

- ramka danych z .csv (duże BoW [Bag of Words]) W ramce jest 8266 kolumn, z których każda przedstawia inne słowo. Każda komórka zawiera liczbę wystąpień danego słowa (kolumny) w wybranym rozdziale (wierszu).
- ramka danych z .csv (duże BoW) po PCA
- nasza ramka danych
- nasza ramka danych po PCA

Postanowiliśmy przetestować następujące modele:

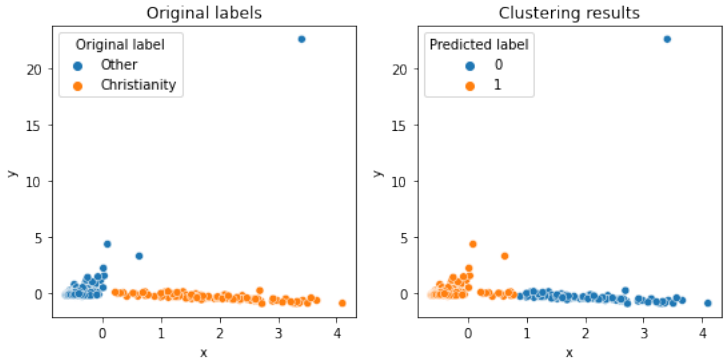
- KMeans
- Agglomerative clustering
- Gaussian mixture model

W celu przyjrzenia się efektywności wypróbowanych modeli, wykorzystaliśmy 3 metryki dostępne w pakiecie sklearn:

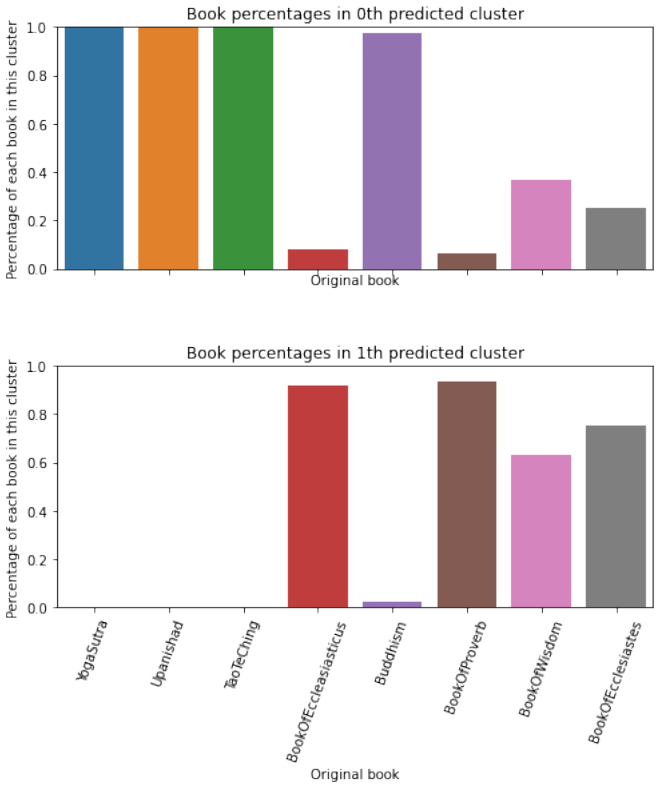
- silhouette_score (im wyższa, tym lepiej, zakres od -1 do 1),
- davies_bouldin_score (im niższa, tym lepiej, zakres od 0 do +nieskończoności),
- calinski_harabasz_score (szukamy szczytów, gładka linia oznacza „brak różnic")

Aby lepiej przedstawić wyniki, stworzyliśmy proste dwuwymiarowe wizualizacje dla każdej z kombinacji zbiór + model.
Dodatkowo dla każdej metody po wybraniu liczby klastrów wygenerowaliśmy wykres pokazujący procent każdej oryginalnej etykiety obecnej w każdym utworzonym skupieniu. Innymi słowy, jest to procent wszystkich rozdziałów z danej książki zgrupowanych w każdym klastrze dla naszych rozwiązań klastrowych.

KMeans
Dla tej metody najbardziej optymalną liczbą klastrów okazało się 2 lub 3.



Rysunek 2: Porównanie wyników klastrowania (prawy wykres) z oryginalnymi etykietami (lewy wykres). Klastrowanie z użyciem KMeans na 2 klastry zbioru naszego (nie przekształcanego PCA).

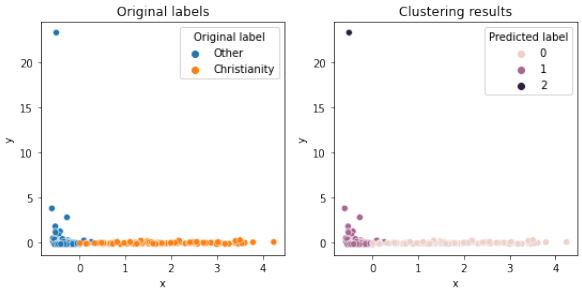


Rysunek 3: Kolejne wykresy przedstawiają w słuphach danego koloru, ile procent rozdziałów z danej książki zostało przyporządkowanych do danego klastra. Rozdziały z pierwszych 3 książek zostały zaklasyfikowane w całości do jednego, a pozostałe nieco bardziej pomieszane są.

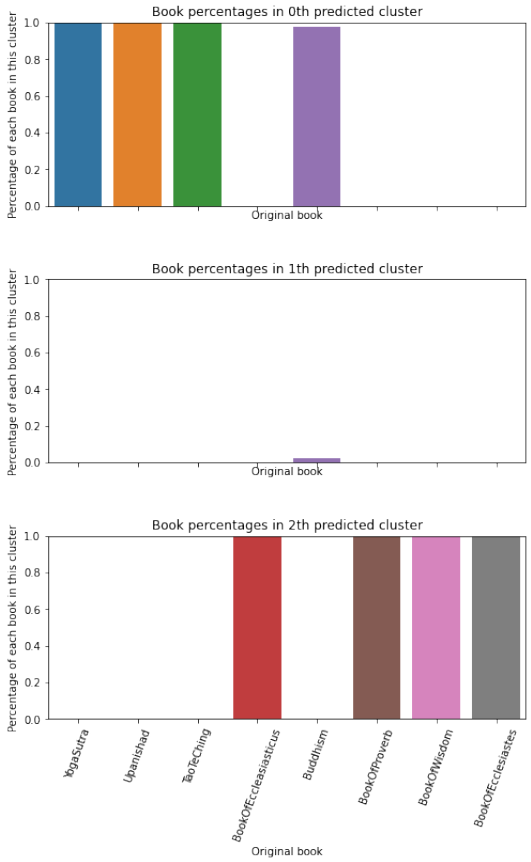
Gaussian mixture model
Przy wykorzystaniu tej metody okazało się, że najlepiej byłoby użyć 9 klastrów, co jest dosyć zaskakujące, biorąc pod uwagę, że w sumie jest 8 kategorii ksiąg (ewentualnie 4 lub 2 po uproszczeniu). Wbrew sugestiom z wykresów mierzących to, jak dobrze są klastry dobrane, liczba 9 jest z czapy i po weryfikacji nie jest sensownym klastrowaniem (większość klastrów jest pusta/bardzo mało liczna).

Agglomerative clustering
Według metryk, najlepszą liczbą klastrów jest 3, jednak okazuje się, że jest to 2:

- jeden z klastrów zawiera wyłącznie rozdziały z książek 1,2,3 i 5 (wg Rysunku 5),
- drugi z klastrów zawiera wyłącznie rozdziały z książek 4, 6, 7 i 8 (wg Rysunku 5),
- trzeci z klastrów zawiera pojedynczy rozdział.



Rysunek 4: Porównanie wyników klastrowania (prawy wykres) z oryginalnymi etykietami (lewy wykres). Klastrowanie z użyciem Agglomerative clustering na 3 klastry oryginalnego zbioru z csv (Bag of Words, ponad 8000 słów, bez przekształceń).



Rysunek 5: Agglomerative clustering - klastry pierwszy i trzeci bardzo dobrze oddzielają od siebie rozdziały z książek chrześcijańskich vs wschodnich. Klaster drugi jest do praktycznie pusty, więc do pominięcia.

4 Wnioski i obserwacje

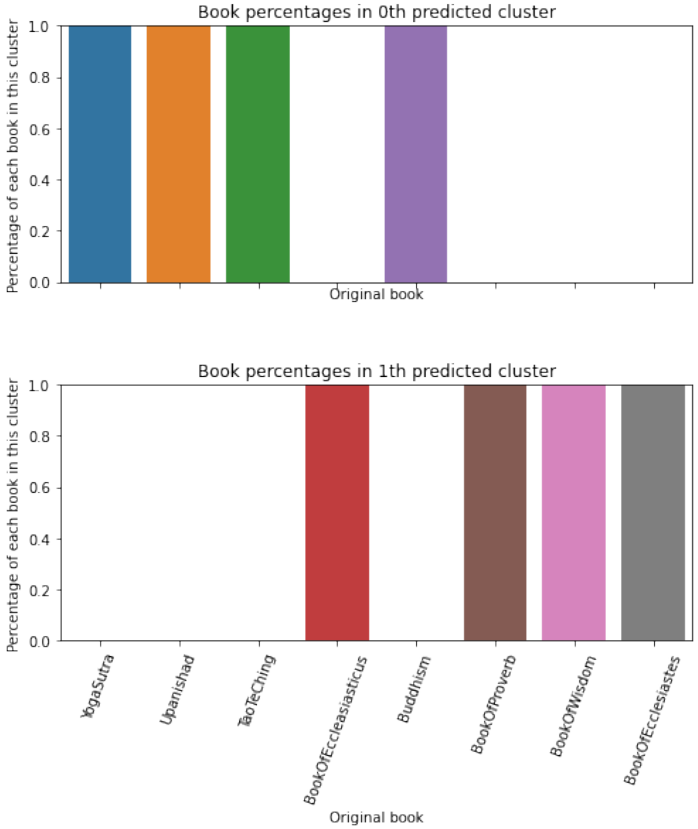
Ogólnie rzecz biorąc, wybór modelu był w tym przypadku równie ważny jak inżynieria cech. Naszym zdaniem, najlepszym sposobem pogrupowania tych 590 rozdziałów w grupy jest użycie KMeans lub klastrowania aglomeracyjnego z 2 klastrami.

Po pracy z tym zbiorem danych dowiedzieliśmy się kilku rzeczy o nienadzorowanym klastrowaniu. Jedną z nich jest to, że PCA drastycznie zmienia wyniki (nie zawsze na lepsze) i znacznie zmniejsza rozmiar danych do pracy.

Co więcej, nawet znajomość pierwotnej liczby skupisk w tym zadaniu nie pomogła w odtworzeniu oryginalnych 8 ksiąg z 589 rozdziałów (jeden rozdział był pusty, dlatego nie braliśmy go pod uwagę). Zamiast tego, najlepsze, co możemy zrobić, to oddzielić rozdziały z religii chrześcijańskiej od tekstów pochodzących z religii wschodnich.

5 Podsumowanie

Z przeprowadzonych eksperymentów wynika, że klasyfikacja najlepiej sprawdza się przy podziale wszystkich danych na 2 klastry. Po sprawdzeniu oryginalnych etykiet okazuje się, że jest to dokładne oddzielenie tekstów z religii wschodnich od tekstów chrześcijańskich, jak jest to przedstawione na Rysunku 6 poniżej.



Rysunek 6: Oddzielenie tekstów chrześcijańskich od tekstów z religii wschodnich bez żadnego błędu. Agglomerative Clustering z 2 klastrami na danych pochodzących prosto z pliku csv (BoW, 8000+ słów) po odrzuceniu pustego rozdziału, bez żadnych przekształceń. Pierwszy klaster zawiera wyłącznie oraz wszystkie rozdziały z ksiąg nie-chrześcijańskich, natomiast klaster drugi zawiera wyłącznie i wszystkie rozdziały z ksiąg chrześcijańskich.