

WUM-projekt2

Agata Kaczmarek

Maria Kałuska

Marcelina Kurek

28.05.2021

1 Zbiór danych

Link do zbioru danych Nasz zbiór składał się z 590 tekstów religijnych z ośmiu ksiąg z czterech religii.

1. Chrześcijaństwo - Księga Madrości, Księga Przysłów, Księga Koheleta * 2
2. Hinduizm - Upanishads, Yoga Sutras,
3. Buddyzm - Buddha Sutras,
4. Taoizm - Tao Te Ching

2 Cel projektu

Projekt dotyczył wykorzystania przetwarzania języka naturalnego w celu klasteryzacji tekstów religijnych. Otrzymane rozwiązanie zakładało podział tekstów na grupy w sposób zbliżony do rzeczywistego podziału ksiąg na religie. W tym celu wykorzystaliśmy ramkę danych *Complete_data.txt* zawierająca pełen dostępny tekst. Postanowiłyśmy porównać wyniki dla ramki danych przekształconej przy pomocy inżynierii cech oraz z wykorzystaniem jedynie zmniejszenia wymiarowości danych przy użyciu PCA.

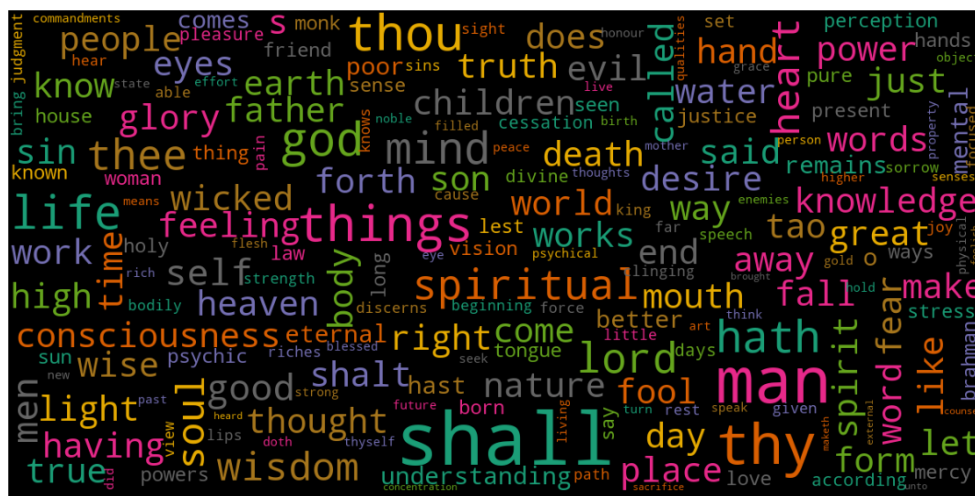


Figure 1: Chmura słów dla wszystkich tekstów religijnych łącznie

3 Inżynieria cech

W ramach określania struktury tekstów, dla każdego z nich wyodrębniłyśmy następujące cechy:

1. Liczba wyrazów,
2. liczba zdań,
3. liczba liter,
4. średnia długość wyrazu,
5. średnia długość zdania,
6. liczba wyrazów bez stopwords,
7. liczba wyrazów bez uwzględnienia powtórzeń.

Kolejnym krokiem była analiza bloku tekstu. W tym celu dodałyśmy dwie cechy:

1. polarność (polarity) - nacechowanie tekstu pozytywne - 1, negatywne - (-1),
2. subiektywność (subjectivity) - określenie czy tekst jest subiektywny - 1, czy obiektywny - 0.

Następnie analizowałyśmy złożoność tekstów. Skorzystaliśmy z dwóch wskaźników FRE (Flesh Reading Ease) oraz ARI (Automated Readability Index). Dość wyraźnie okazało się, że teksty religijne należą do tekstów trudnych. Według FRE teksty zostały uznane za nieznacznie trudne, z kolei ARI wskazywał, że aby wpłynąć zrozumieć tekst powinno się mieć wykształcenie profesora.

Kolejnymi cechami, które postanowiłyśmy określić była tematyka tekstu. Pierwszym podejściem było skorzystanie z algorytmu LDA na worku słów (Bag of Words), który generował osiem zbiorów słów, które najczęściej występowały wśród tekstów. Drugim sposobem na wyznaczenie tematów było wpięrow przypisanie wag słowom dla każdego z tekstów, a następnie zastosowanie na tak przygotowanej ramce danych algorytmu NMF(Non Negative Matrix Factorization), który podobnie jak LDA wyznacza najwyraźniej przebiegające się tematy.

3.1 Kodowanie i normalizacja zmiennych

Wszystkie cechy kategoryczne zostały zakodowane przy pomocy One Hot Encodera, z kolei wszystkie zmienne numeryczne wykraczające poza zakres $[-1, 1]$ zostały zestandaryzowane.

4 Modelowanie

4.1 AgglomerativeClustering

Jednym z wykorzystanych modeli była klasteryzacja aglomeracyjna. Początkowo sprawdziliśmy optymalną liczbę klastrów z wykorzystaniem indeksu silhouette, davisa-bouldina oraz calińskiego-harabasa. W przypadku wszystkich metryk optymalny podział zakładał od 2 do 4 klastrów. Z kolei dendrogram dla modelu aglomeracyjnej klasteryzacji wskazywał 2 lub 3 klastry jako adekwatną liczbę. Wykorzystując kolumnę *Labels* porównaliśmy wartości indeksów fowlkesa mallows oraz completeness. Najwyższe wyniki metryk uzyskał model z liczbą klastrów równą 2 - indeks fowlkesa mallows równy 0.77 oraz completeness na poziomie 0.94. Następnie porównaliśmy wyniki z modelami o zredukowanej wymiarowości przy pomocy PCA. Transformacja przy użyciu PCA spowodowała polepszenie indeksu silhouette, davisa-bouldina i calińskiego-harabasa oraz pogorszenie indeksu completeness średnio o 0.11. Wartości pozostałych metryk są porównywalne przed i po zastosowaniu algorytmu PCA.

4.2 KMeans

W przypadku tej metody klasteryzacji proponowana liczba klastrów również wynosiła 2 lub 3, dodatkowo postanowiliśmy uwzględnić w dalszym testowaniu podział na 4 klastry. Metode KMeans zastosowaliśmy na przygotowanych przez nas danych jak również na danych o zredukowanej wymiarowości przy pomocy PCA. Otrzymane wyniki wyszły bardzo podobne.

4.3 Alternatywne podejście

Na koniec postanowiliśmy sprawdzić co by się stało, gdybyśmy nie przeprowadziły powyżej opisanej skomplikowanej inżynierii cech, a tylko zredukowały wymiary danych przy pomocy PCA. Następnie również przeprowadziłyśmy klastrowanie, przy użyciu metody KMeans. Również i tym razem proponowana liczba klastrów wynosiła 2 lub 3, także dodałyśmy podział na 4 klastry. Zależnie od liczby klastrów otrzymane wyniki tylko nieznacznie różniły się od tych z poprzednich metod (dla 2 klastrów) jednak dla 4 klastrów różnice w wynikach były już znaczne.

5 Podsumowanie

Wybrałyśmy trzy metryki według których porównałyśmy końcowo otrzymane wyniki: accuracy score, fowlkes mallows score oraz completeness score. Najmniejsze różnice między zastosowanymi metodami były zauważalne przy podziale na dwa klastry.

	accuracy score	fowlkes mallows score	completeness score
KMeans 2	0.774576	0.762999	0.834355
KMeans 2PCA 2	0.774576	0.764587	0.824334
AgglomerativeClustering 2	0.783051	0.772234	0.948163
AgglomerativeClustering 2PCA 2	0.774576	0.763783	0.829040
KMeans 2 alternative 2	0.764407	0.746148	0.842345

Figure 2: Podział na dwa klastry

Jednak nie to było głównym celem projektu - wiemy, że nasz zbiór zawierał cztery religie, więc skuteczność podziału na cztery klastry chcielibyśmy również porównać. I w tym wypadku już zauważyliśmy znaczne różnice. Po pierwsze można było zauważyć znaczny spadek wartości dla jednej z metryk (completeness score). Po drugie pojawiła się metoda, która wypadła dla większości metryk znacznie lepiej niż pozostałe - Agglomerative Clustering - wersja bez zastosowania PCA.

	accuracy score	fowlkes mallows score	completeness score
KMeans 4	0.827119	0.583532	0.443573
KMeans 4PCA 2	0.803390	0.725487	0.513978
AgglomerativeClustering 4	0.803390	0.730973	0.669681
AgglomerativeClustering 4PCA 2	0.815254	0.730888	0.513247
KMeans 4 alternative 4	0.764407	0.531763	0.435159

Figure 3: Podział na cztery klastry

Warto również porównać osobno wyniki uzyskane przez KMeans dla różnie przerobionych danych - z użyciem inżynierii cech oraz bez jej użycia. Jak widać w poniższej tabeli, inżynieria cech zastosowana dla ramki danych skutkowała polepszeniem jakości klasteryzacji. Przykładowo wartość metryki accuracy (podział na cztery klastry) z jej zastosowaniem wyniosła 0.82 oraz 0.76 dla ramki przekształconej przy pomocy PCA (bez inżynierii cech).

	accuracy score	fowlkes mallows score	completeness score
KMeans 2	0.774576	0.762999	0.834355
KMeans 2 alternative 2	0.764407	0.746148	0.842345
KMeans 4	0.827119	0.583532	0.443573
KMeans 4 alternative 4	0.764407	0.531763	0.435159

Figure 4: Porównanie użycia inżynierii cech