1 Zbiór danych

Link do zbioru danych Nasz zbiór składał sie z 590 tekstów religijnych z ośmiu ksiag z czterech religii.

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

2 Cel projektu

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

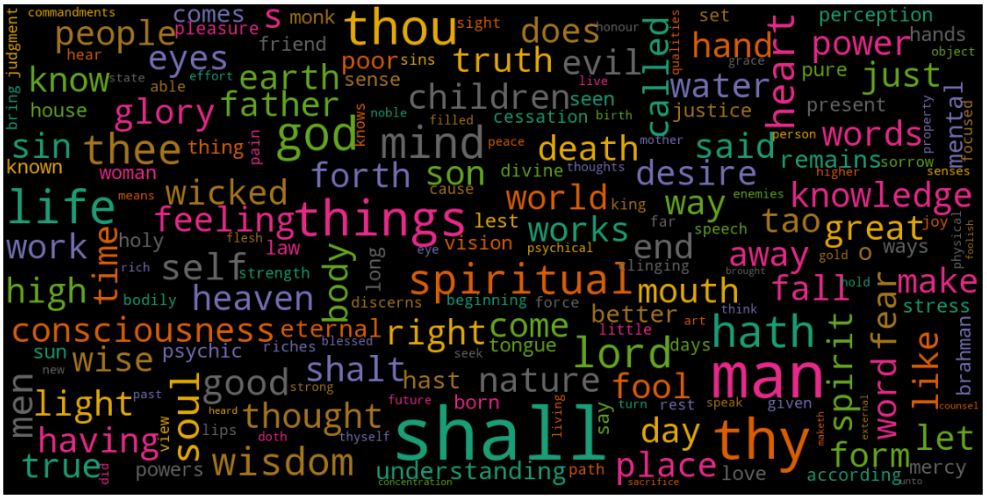

Figure 1: Chmura słów dla wszystkich tekstów religijnych łacznie

3 Inżynieria cech

W ramach określania struktury tesktów, dla każdego z nich wyodrebniłyśmy nastepujace 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 uwzglednienia 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.

Nastepnie analizowałyśmy złożoność tekstów. Skorzystałyśmy z dwóch wskaźników FRE (Flesh Reading Ease) oraz ARI (Automated Readibility Index). Dość wyraźnie okazało sie, że teksty religijne należa do tekstów trudnych. Według FRE teksty zostały uznane za nieznacznie trudne, z kolei ARI wskazywał, że aby wpełni zrozumieć tekst powinno sie mieć wykształcenie profesora.
Kolejnymi cechami, które postanowiłyśmy określić była tematyka tesktu. Pierwszym podejściem było skorzystanie z algorytmu LDA na worku słów (Bag of Words), który generował osiem zbiorów słów, które najcześciej wystepowały wśród tekstów. Drugim sposobem na wyznaczenie tematów było wpierw przypisanie wag słowom dla każdego z tekstów, a nastepnie zastosowanie na tak przygotowanej ramce danych algorytmu NMF(Non Negative Matrix Factorization), który podobnie jak LDA wyznacza najwyraźniej przebijajace sie tematy.

3.1 Kodowanie i normalizacja zmiennych

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

4 Modelowanie

4.1 AgglomerativeClustering

Jednym z wykorzystanych modeli była klasteryzacja aglomeracyjna. Poczatkowo sprawdziłyśmy optymalna liczbe klastrów z wykorzystaniem indeksu silhouette, davisa-bouldina oraz calińskiego-harabasza. 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 adekwatna liczbe. Wykorzystujac kolumne *Labels* porównałyśmy wartości indeksów fowlkess mallows oraz completeness. Najwyższe wyniki metryk uzyskał model z liczba klastrów równa 2 - indeks fowlkess mallows równy 0.77 oraz campleteness na poziomie 0.94. Nastepnie porównałyśmy wyniki z modelami o zredukowanej wymiarowości przy pomocy PCA. Transformacja przy użyciu PCA spowodowała polepsznie indeksu silhouette, daviesa-bouldina i calinskiego-harabasza oraz pogorszenie indeksu completeness średnio o 0.11 Wartości pozostałych metryk sa porównywalne przed i po zastosowaniu algorytmu PCA.

4.2 KMeans

W przypadku tej metody klasteryzacji proponowana liczba klastrów również wynosiła 2 lub 3, dodatkowo postanowiłysmy uwzglednić w dalszym testowaniu podział na 4 klastry. Metode KMeans zastosowałyśmy 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 postanowiłyśmy sprawdzić co by sie stało, gdybyśmy nie przeprowadziły powyżej opisanej skomplikowanej inżynierii cech, a tylko zredukowały wymiary danych przy pomocy PCA. Nastepnie 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 sie 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 miedzy 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, wiec skuteczność podziału na cztery klastry chciałybyśmy również porównać. I w tym wypadku już zauważyłyśmy znaczne różnice. Po pierwsze można było zauważyć znaczny spadek wartości dla jednej z metryk (completeness score). Po drugie pojawiła sie metoda, która wypadła dla wiekszości 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