Wstęp do uczenia maszynowego 2020

PROJEKT 2.
PROBLEM KLASTERYZACJI DLA ZBIORU *ASIAN RELIGIONS DATA*

8 czerwca 2020

Jan Borowski, Elżbieta Jowik, Marceli Korbin

1 Opis problemu i rozważanego zbioru danych

Klastrowanie jest jedną z najczęściej stosowanych technik eksploracyjnej analizy danych. Ma ono na celu wykrycie w zbiorze obserwacji skupień, czyli **rozłącznych** podzbiorów zbioru, wewnątrz których obserwacje są sobie bliskie względem określonego kryterium, natomiast różne podzbiory są od siebie odległe. Innymi słowy celem analizy skupień jest znalezienie w zbiorze danych ukrytej struktury.
Przedmiotem naszych analiz, pod kątem obecności owej struktury, jest zbiór danych ***Asian Religions***. Zawiera on wyselekcjonowany przez twórców podzbiór wyrazów występujących w księgach religijnych. Większość tekstów, na podstawie których opracowano zbiór pochodzi z projektu *Gutenberg*. Uwzględniono 8 ksiąg, w badaniu rozważono 8265 słów.
Zbiór ma problematyczną, z punktu widzenia klasteryzacji, strukturę. Zauważalna jest bowiem nietypowa dysproporcja pomiędzy licznościami kolumn i wierszy. W zbiorze występuje 8265 atrybutów i jest ich znacznie więcej niż rekordów, których liczność wynosi 590. W rozważanym zbiorze wiersze odpowiadają kolejnym rozdziałom poszczególnych ksiąg, natomiast kolumny są binarnymi *wektorami*, w których *1* jest tożsama z obecnością, a *0* z brakiem obecności danego słowa w określonym przez wiersz rozdziale. Poza pierwszą zmienną objaśniającą, tożsamą z nazwami rozdziałów poszczególnych ksiąg, wszystkie pozostałe mają charakter liczbowy. Ponadto, zbiór nie zawiera żadnych brakujących danych.

2 Eksploracyjna analiza danych

Przeprowadzona analiza eksploracyjna dowiodła, że spośród wszystkich 590 rozdziałów ponad 38% zawiera poniżej 50 słów, a odsetek wyrazów unikalnych w zbiorze to niespełna 43%. Ponadto, w toku eksploracji okazało się, że 99.14% danych w zbiorze stanowią zera.
Powyższe obserwacje uzasadnieniają fakt występowania dużej liczby par silnie skorelowanych wyrazów.
Zbiór danych przeanalizowaliśmy również pod kątem obecności synonimów, słów pochodzących z tych samych rodzin wyrazów oraz często używanych części zdania, takich jak spójniki, czy zaimki. W rezultacie mogliśmy założyć, że w ogólności w zbiorze niezauważalny jest problem występowania wyrazów bliskoznacznych, czy tych najpopularniejszych w mowie i piśmie. Natomiast obecność wyrazów pokrewnych jest ewidentna i stanowi jedną z przyczyn konieczności przeprowadzenia kompleksowego czyszczenia zbioru danych.

3 Czyszczenie danych

W czyszczeniu danych nie trzeba było uwzględniać modyfikacji typów zmiennych. Istniały natomiast liczne kolumny, które prawdopodobnie powstały przypadkowo (w wyniku błędów w przetwarzaniu tekstów), albo które są wspomnianymi wyrazami pokrewnymi. Przy oddzielaniu tych pierwszych słów liczyliśmy ich wystąpienia jako obecności rzeczywistych słów wewnątrz. Etapy czyszczenia:

- usunięcie kolumn "TRUE" i "FALSE";
- rozdzielenie słów dwukrotnych;
- oddzielenie "consciusness" na końcu;
- rozdzielenie słów o schemacie "neither...nor...";
- ręczne rozdzielenie takich słów, które dokładnie w jednym miejscu dzielą się na dwa inne, rozpoznawane przez pakiet Wordnet (będący częścią NLTK), przeznaczony do analizy języka i słownictwa;
- połączenie słów, które są odmianą któregoś krótszego, z oryginalną formą;
- połączenie czasowników o staroangielskich końcówkach koniugacyjnych "eth" i "est";
- ręczne oddzielenie kilku innych kolumn;
- połączenie słów, których Wordnet nie rozpoznaje, a które stanowią formę rozpoznawanego słowa połączoną z przyrostkiem typu "ation", "ism", "ness", "ship".

Po wyczyszczeniu danych, liczba kolumn spadła do 4806 (58% poprzedniej liczby).

4 Inżynieria cech & wstępne modelowanie

W celu zwiększenia mocy predykcyjnej algorytmów uczenia maszynowego, przed przystąpieniem do modelowania należy odpowiednio przygotować charakterystyki zbioru danych. W naszym przypadku w procesie inżynierii cech można wyodrębnić dwa, opisane poniżej, etapy: zanurzenie (ang. *embedding*) i redukcję wymiarów.

4.1 Embedding

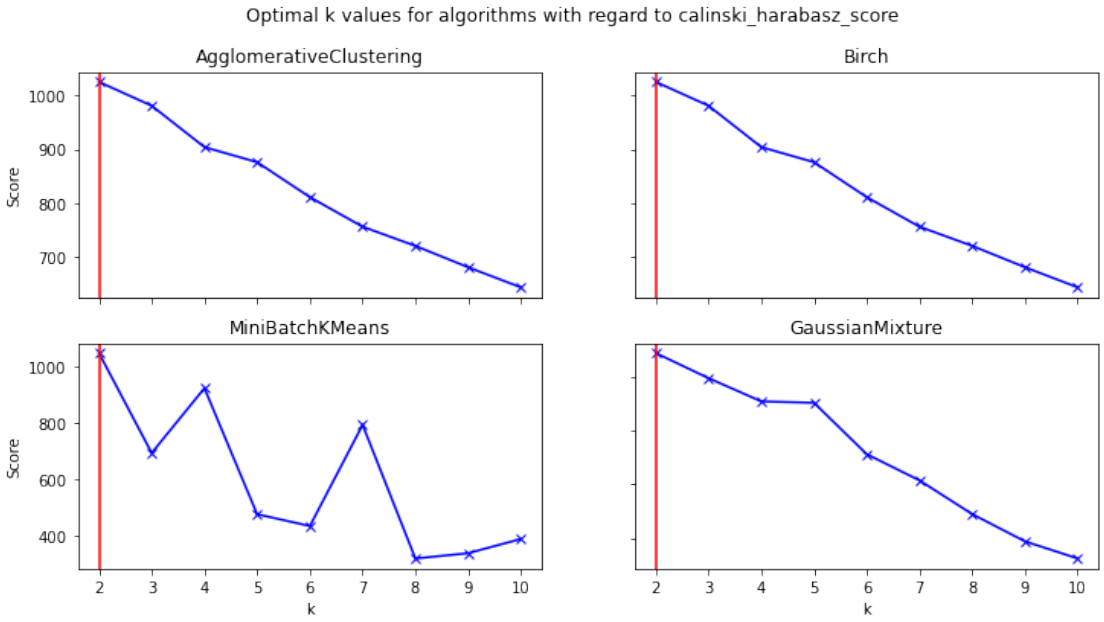
Zdecydowaliśmy się zastosować embedding *Glove* zainplementowany w pakiecie *flair*. W poznaniu pełnej specyfiki pakietu i wybranego embeddingu pomocne będą źródła odpowiednio:
`https://github.com/flairNLP/flair/tree/master/resources/docs`
`https://nlp.stanford.edu/projects/glove/`
Embedding to przekształcenie przypisujące słowu w naszym przypadku stuelementowy

wektor liczbowy. Kluczową cechą embeddingów jest fakt, że bliskoznaczne słowa znajdują się blisko w tej przestrzeni. Istotna dla naszych celów jest ich addytywność, tj. embeddingi dla słowa *król* plus różnica embeddingów słów *mężczyzna*, *kobieta* da wynik bardzo bliski słowa *królowa*. Własność ta jest bardzo istotna w bieżących rozważaniach, ponieważ analizowany problem wymaga reprezentacji rozdziałów, a nie poszczególnych słów. Początkowo zdecydowaliśmy się utworzyć cztery ramki danych. Każda obserwacja w tych ramkach reprezentowała jeden rozdział:

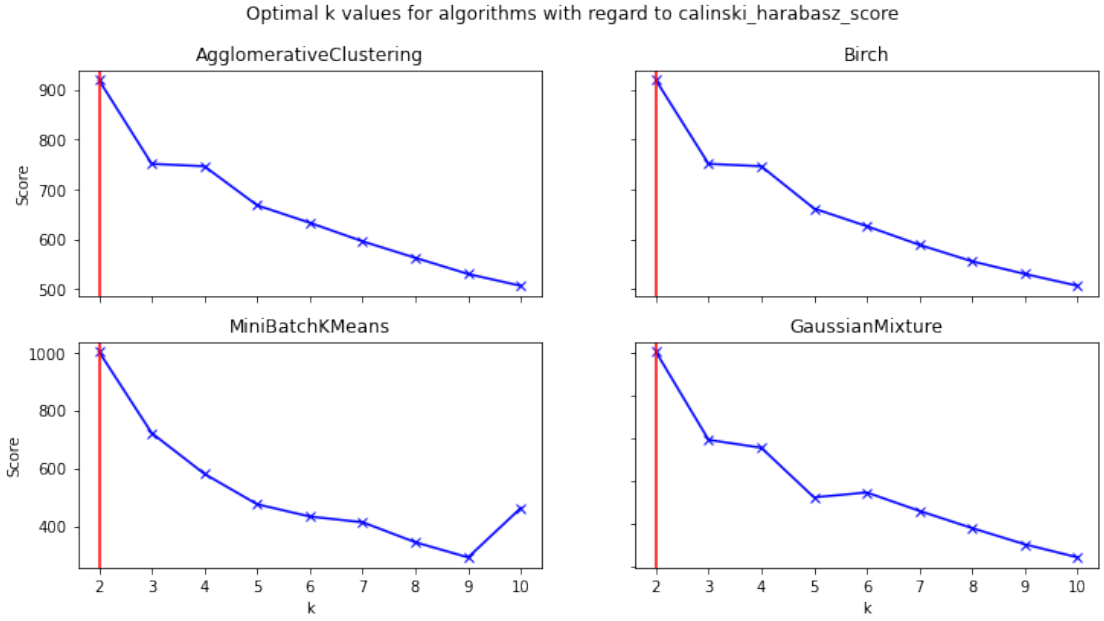
- **Ramka 1**: Wektor reprezentujący każdy embedding to suma ważona embeddingów słów występujących w danym rozdziale, gdzie wagą jest ilość wystąpień słowa w rozdziale.
- **Ramka 2**: Rozumowanie analogiczne do powyższego, przy czym wagi zostały poddane logarytmizacji.
- **Ramka 3**: Również suma ważona, ale tym razem wagą jest ilość wystąpień słowa podzielona przez liczbę słów w rozdziale. Takie działanie ma na celu utrzymanie blisko siebie podobnych rozdziałów o różnej długości.
- **Ramka 4**: Suma embeddingów wszystkich słów występujących w rozdziale bez nadawania wag.

Wstępna weryfikacja dla przygotowanych ramek dowiodła, że najlepsze wyniki uzyskujemy wykorzystując ramek: 1. i 4. i to na ich analizę zorientowana była dalsza praca.
W tym miejscu przytoczę rezultaty otrzymane względem pierwszej spośród wymienionych powyżej metryk:

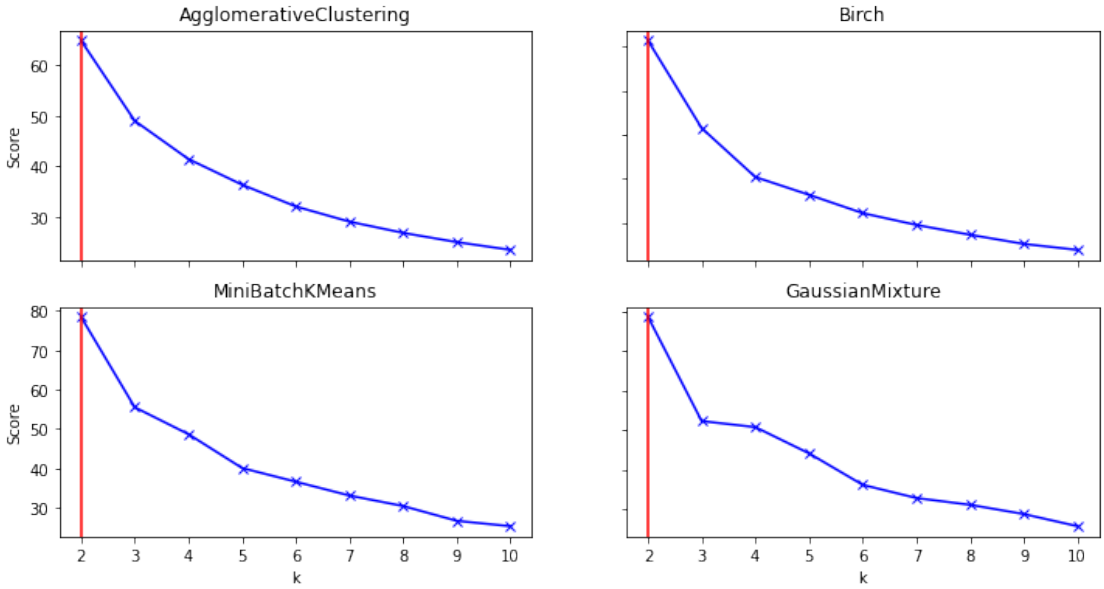
Rysunek 1: Wyniki CH dla Ramki 1



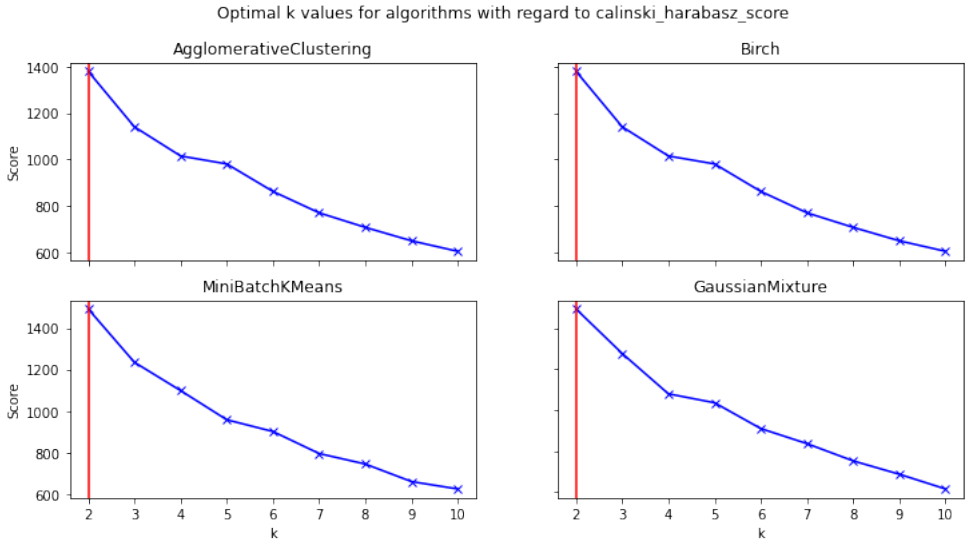
Rysunek 2: Wyniki CH dla Ramki 2



Rysunek 3: Wyniki CH dla Ramki 3



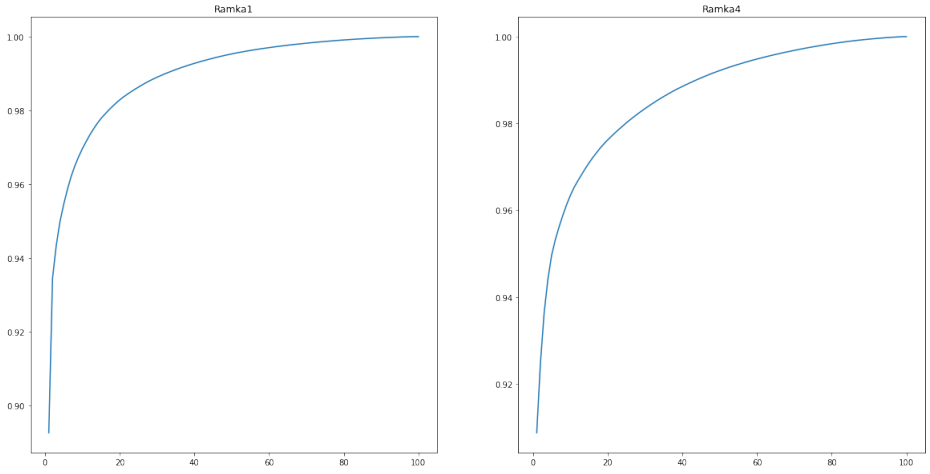
Rysunek 4: Wyniki CH dla Ramki 4



4.2 Redukcja wymiarów

Przygotowane ramki zdecydowaliśmy się dodatkowo przetworzyć, wykorzystując do tego dwa algorytmy redukcji wymiarów: analizę głównych składowych (ang. *Principal Component Analysis* (PCA)) oraz stochastyczną metodę porządkowania sąsiadów w oparciu o rozkład t (ang. *T-distributed Stochastic Neighbor Embedding* (t-SNE)). Na potrzeby testów wykorzystano dwie miary: indeks Calińskiego-Harabasza oraz indeks Daviesa-Bouldina. Kolejnym krokiem metodologii był wybór odpowiedniej ilości komponentów dla algorytmu PCA.

Rysunek 5: Odsetek wyjaśnionej wariancji w zależności od liczby komponentów

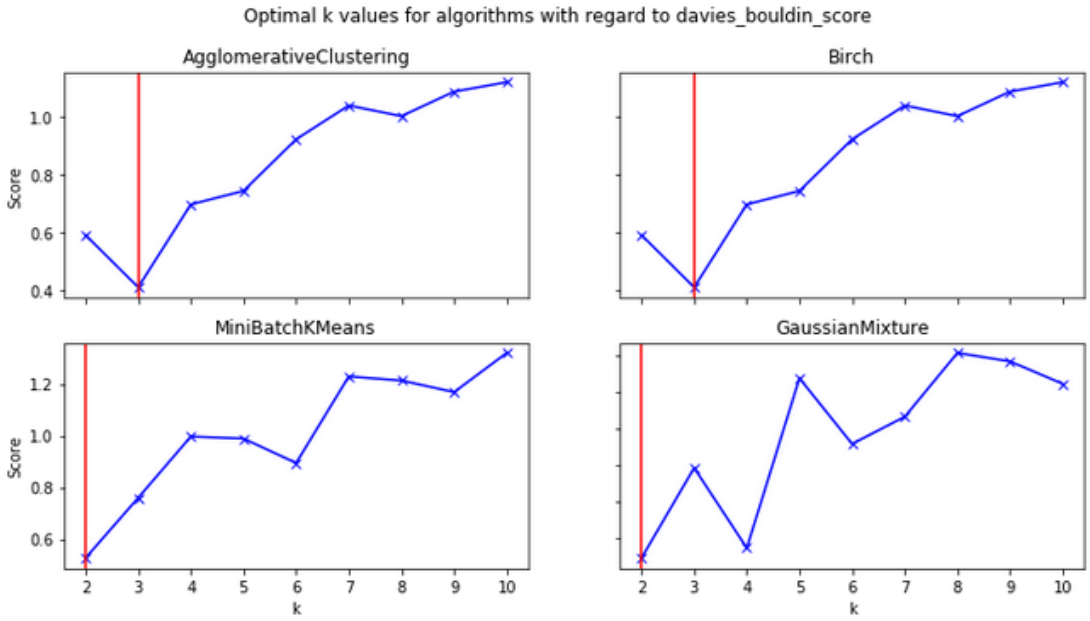


Po analizie powyższego wykresu uznaliśmy, że właściwą liczba składowych wynosi około 20 dla Ramki 1. i około 40 dla Ramki 4. Są to miejsca na wykresie, gdzie zwiększanie liczby komponentów powoduje już tylko marginalny wzrost wyjaśnionej wariancji.
W przypadku t-SNE, analogiczna analiza nie była potrzebna i bez wstępnego szacowania przygotowaliśmy dwie ramki przetworzone za pomocą tego algorytmu.

5 Modelowanie & efekty klasteryzacji

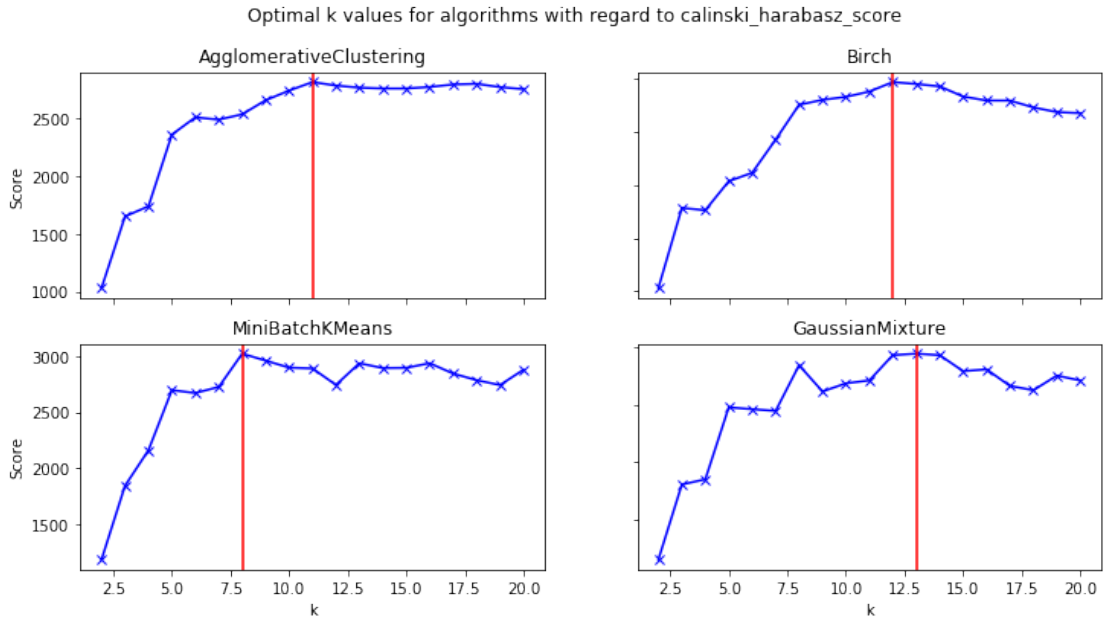
Na skutek przeprowadzonej inżynierii cech, dysponowaliśmy czterema ramkami danych, spośród których dwie przetworzone zostały przez PCA i dwie przez t-SNE. Poniżej najlepsze rezultaty, jakie uzyskaliśmy:

Rysunek 6: Ramka 1 po PCA Indeks DB



Jak wynika z wizualizacji wyniki dla ramek po analizie głównych składowych nie różniły się od wyników dla ramek wyjściowych. Tym nie mniej względem metryki DB *Agglomerative Clustering* i *Birch* dla $k = 3$ pozostały najlepsze.
Sytuacja zmieniła się diametralnie po zastosowaniu t-SNE. Jak wynika z poniższego wykresu, wartości indeksu CH otrzymane po zastosowaniu t-SNE znacznie się zwiększyły. Ponadto, zmieniły się sugerowane optymalne liczby klastrów,

Rysunek 7: Ramka 1 po t-SNE Indeks CH



Po dodatkowej weryfikacji dokładnych rezultatów poszczególnych kombinacji embedding-model okazało się, że najlepsze rezultaty względem indeksu Daviesa-Bouldina uzyskaliśmy dla par:

- Embedding 1. + Agglomerative Clustering, gdzie k = 3,
- Embedding 1. + Birch, gdzie k = 3

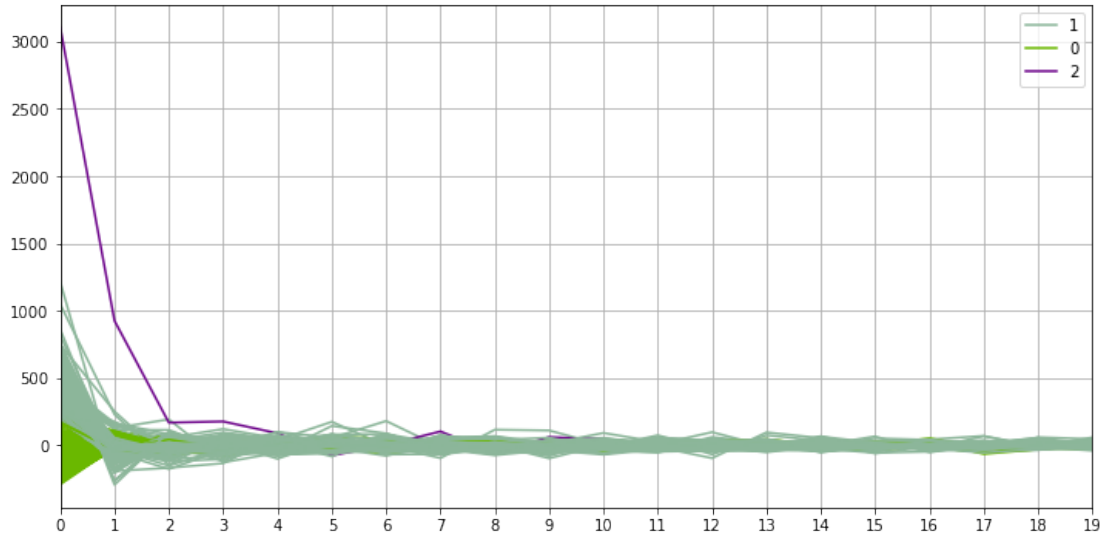
Natomiast względem indeksu Calińskiego-Harabasza dla:

- Embedding 1. + Birch, gdzie k = 12
- Embedding 1. + MiniBatchKMeans, gdzie k = 8

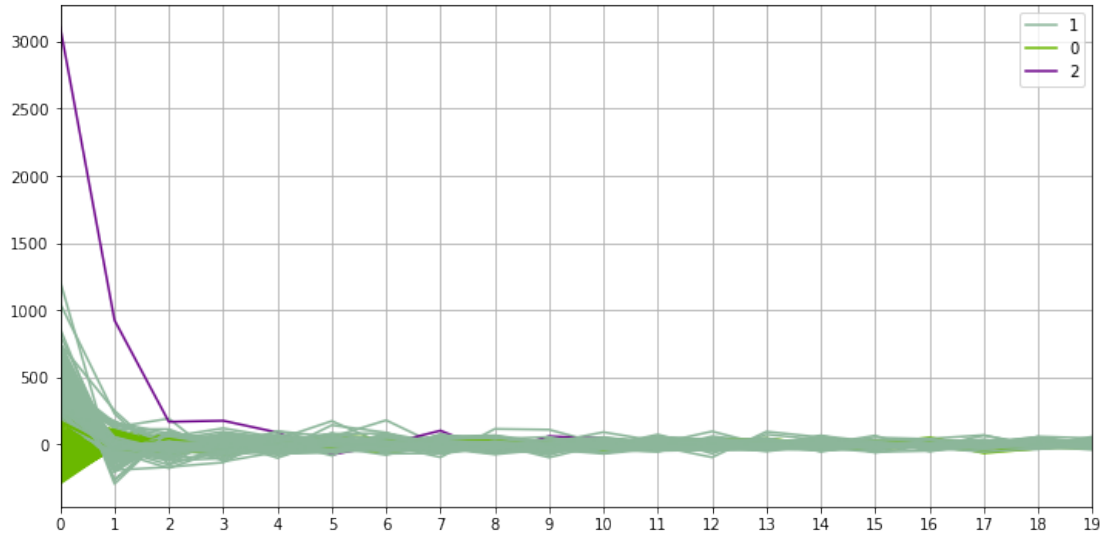
Ponieważ były to globalnie najlepsze kombinacje, spośród których trudno było wyłonić tę optymalną, ostateczne modelowanie przeprowadziliśmy dla każdej z nich.

6 Wizualizacje klasteryzacji

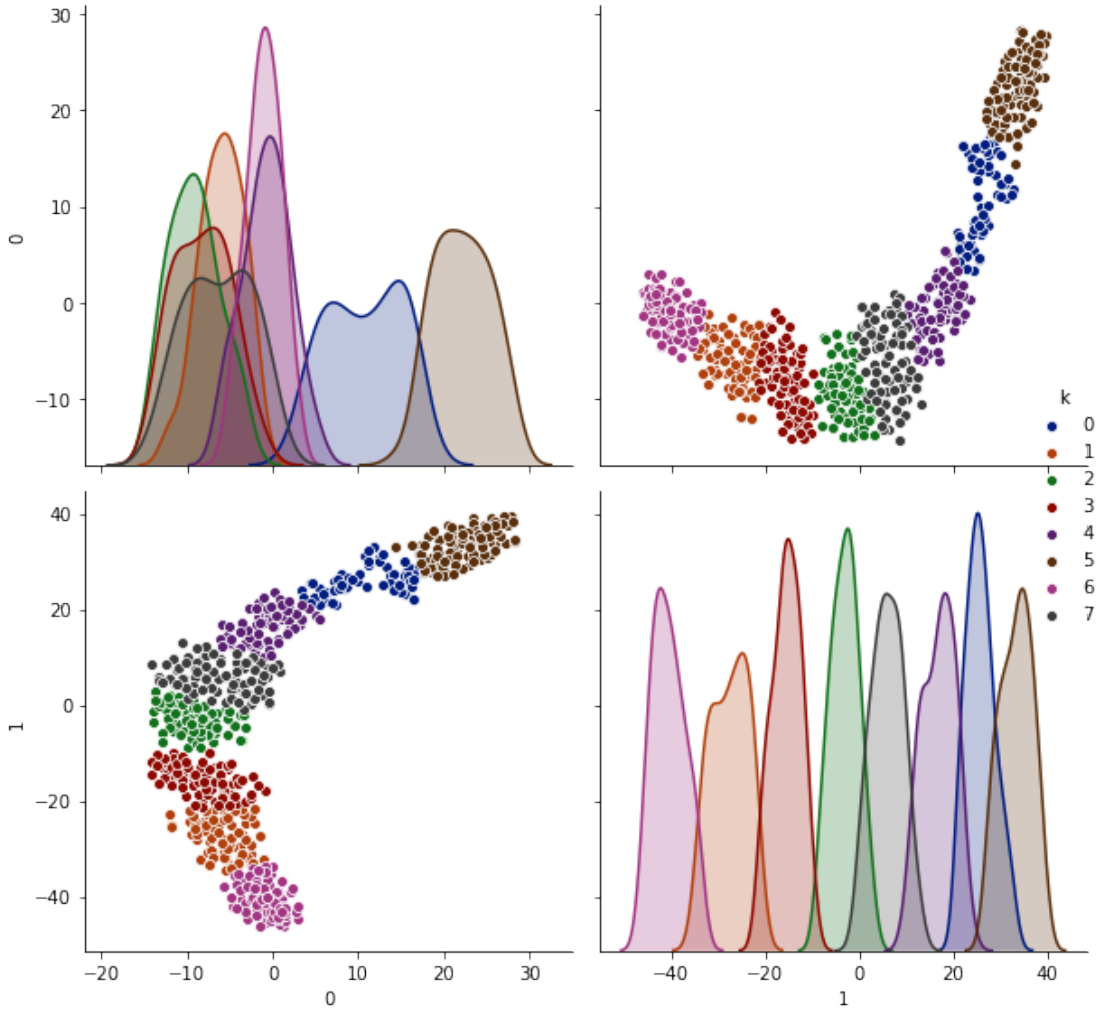
Rysunek 8: Klasteryzacja AgglomerativeClustering dla 3 skupień



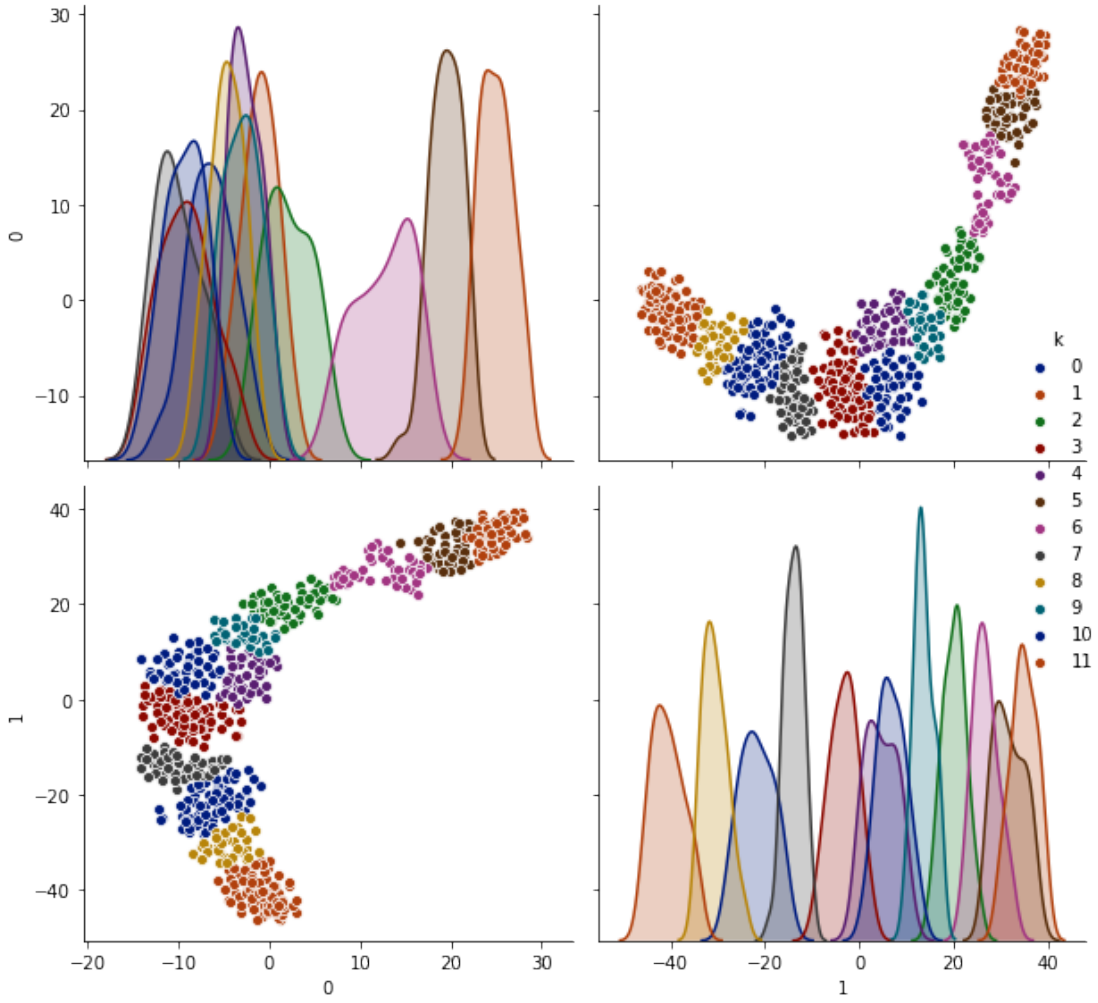
Rysunek 9: Klasteryzacja Birch dla 3 skupień



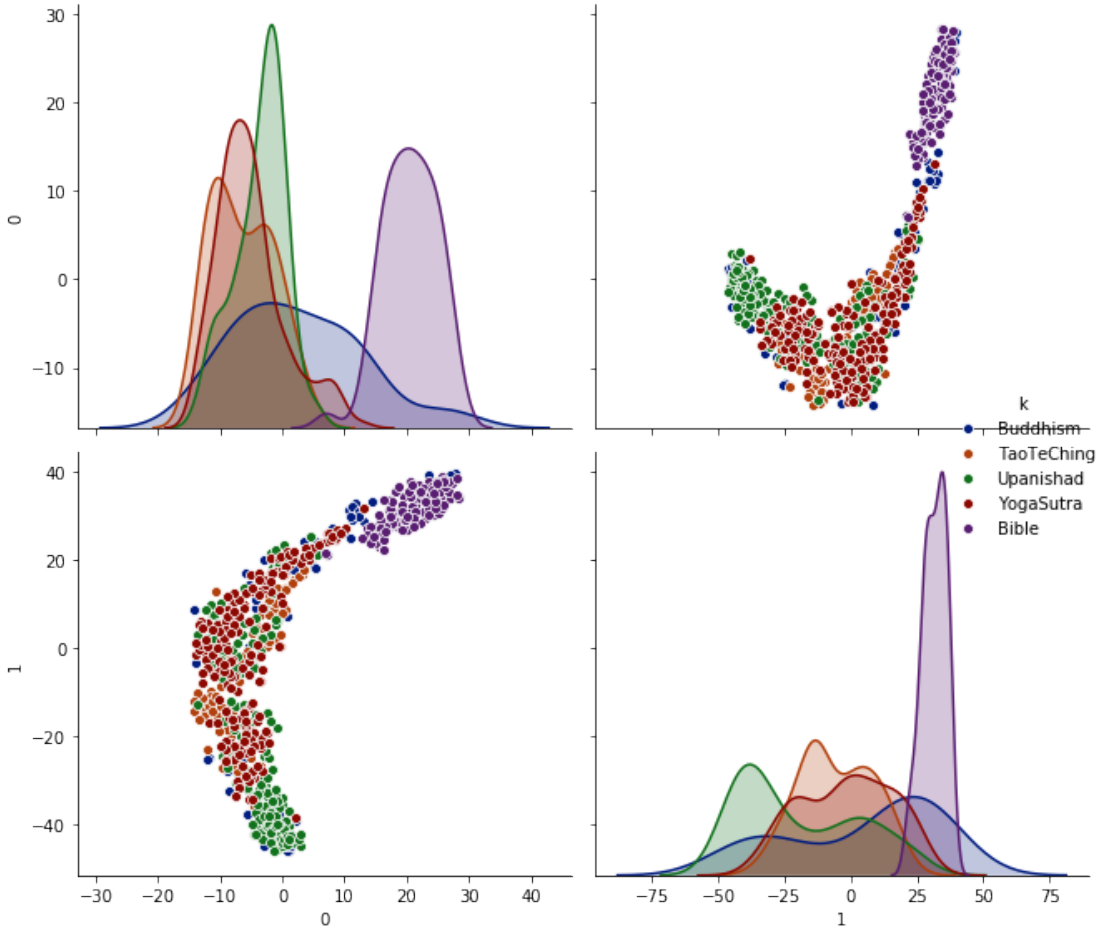
Rysunek 10: Klasteryzacja MiniBatchKMeans dla 8 skupień



Rysunek 11: Klasteryzacja Birch dla 12 skupień



Rysunek 12: Klasteryzacja oryginalna



7 Rezultaty

Istotnym aspektem z punktu widzenia rozważanego problemu, pozwalającym na ocenę uzyskanych skupień jest to jak uzyskane w wyniku klasteryzacji struktury korespondują z faktycznym podziałem na rodziały. W celu realizacji zasygnalizowanego powyższego zagadnienia zbadaliśmy **homogeniczność** przewidywanej struktury z oryginalną, za pomocą metryki homogeneity_score.
Klasteryzacja osiąga wysoką homogeniczność, jeżeli każdy jej podział zawiera tylko obserwacje należące do pojedynczej klasy. Metryka ta jest niezależna od bezwzględnych wartości etykiet: permutacja klasy lub etykiety klastra nie wpływa w żaden sposób na wynik.
Otrzymane pomiary są dalekie od zadowalających. Klasteryzacje uzyskane po metodzie redukcji wymiarów t-SNE były w 40% homogeniczne z oryginalnym podziałem na klastry, natomiast te przekształcone przez PCA tylko w niespełna 30%.

Agglom. & PCA	Birch & PCA	MBKMeans & t-SNE	Birch & t-SNE	Losowe etykiety
0.27440	0.27440	0.40495	0.40495	0.01439

Tabela 1: Przybliżone pomiary homogeniczności podziałów.

Warto jednak zwrócić uwagę na ostatnią z wymienionych miar homogeniczności, jaką sprawdziliśmy dla losowo wygenerowanego podziału (na tę samą liczbę klastrów, co oryginalna). Zrobiliśmy to, aby móc ocenić względną skuteczność naszych klasteryzacji. Miara ta wyniosła 0.01439, co wskazuje, że losowy klastering tylko w 1% jest homogeniczny z zamierzonym. Wobec takiej wartości porównawczej podjęte przez nas działania wydają się mieć większy sens.

8 Wnioski

Pomimo wszelkich starań, maszynowe przyporządkowanie rozdziałów do książek na bazie obecności konkretnych słów nie spełniło wszystkich naszych oczekiwań. Skuteczność najbardziej optymalnych klasteryzacji, wyznaczanych drogą różnorakich eliminacji, okazała się częściowo celna, aczkolwiek daleka od wysokiej.