ŚWIĘTE TEKSTY

Hanna Zdulska, Jakub Wiśniewski

Zbiór danych

W tym zbiorze danych znajdują się następujące książki z bliskiego/ dalekiego wschodu:

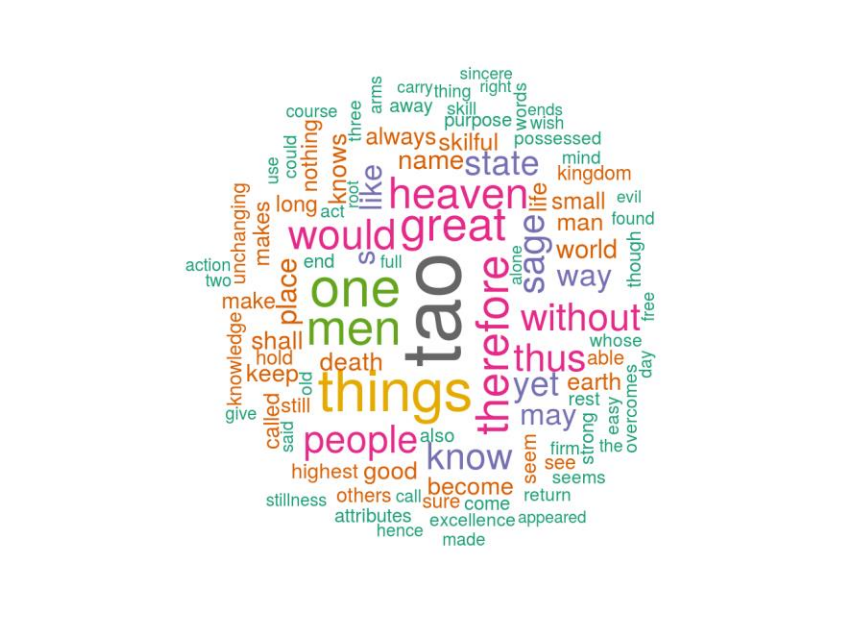
- **Upaniszady** - to starożytne teksty nauk duchowych, mówiące idei hinduizmu. Są one częścią najstarszych pism hinduizmu, Wedy, które dotyczą medytacji, filozofii i wiedzy duchowej; inne części Wedów zajmują się mantrami, błogosławieństwami, rytuałami, ceremoniami i ofiarami.
- **Yoga Sutras** - zbiór 196 sutr sanskryckich (aforyzmów) na temat teorii i praktyki jogi.
- **Budda Sutry** - początkowo były przekazywane ustnie przez mnichów, ale później zostały spisane i skomponowane jako rękopisy w różnych językach indo-aryjskich, które następnie zostały przetłumaczone na inne języki lokalne w miarę rozprzestrzeniania się buddyzmu
- **Tao Te Ching** - podstawowy tekst dla filozoficznego i religijnego taoizmu. Wpłynął silnie na inne szkoły chińskiej filozofii i religii, w tym konfucjanizm i buddyzm, które zostały w dużej mierze zinterpretowane poprzez użycie taoistycznych słów i pojęć, kiedy zostały pierwotnie wprowadzone do Chin.

Zbiór danych

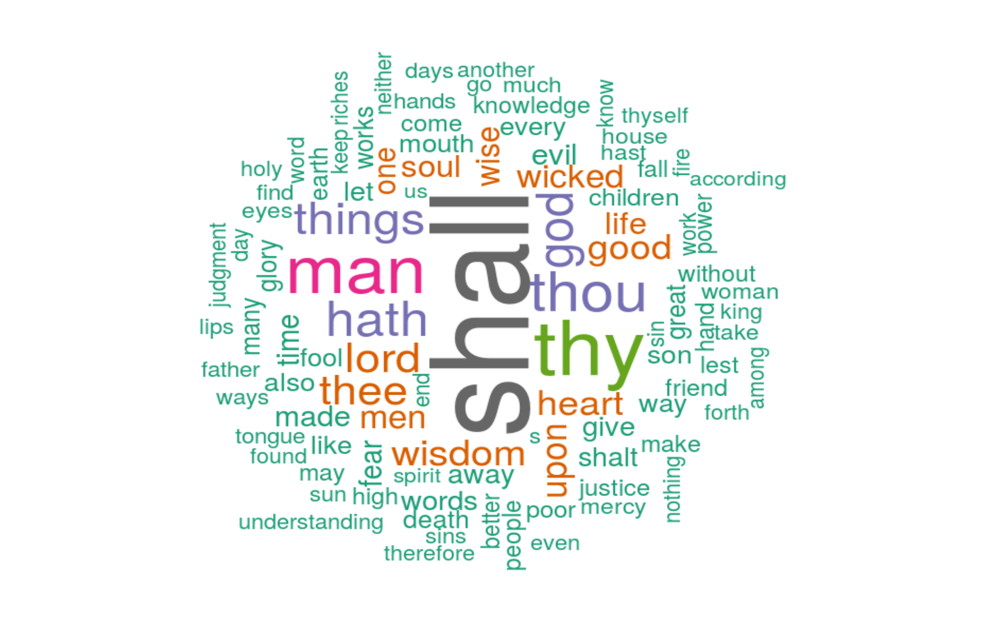
- Oraz Stary Testament :
  - **Księga Mądrości** - mowa Salomona dotycząca mądrości, bogactwa, władzy i modlitwy
  - **Księga Przysłów** - Przysłowia to nie tylko antologia, ale „zbiór kolekcji" odnoszący się do wzoru życia, który trwał ponad tysiąc lat. Jest przykładem biblijnej tradycji mądrościowej i stawia pytania o wartości, moralne zachowanie, sens życia ludzkiego i właściwe postępowanie.
  - **Księga Koheleta** - jest jedną z 24 ksiąg Tanach (hebrajska Biblia).
  - **Mądrość Syracha**

Częstość występowania słów - im większe, tym słowo częściej występuje.

TeoTeChing



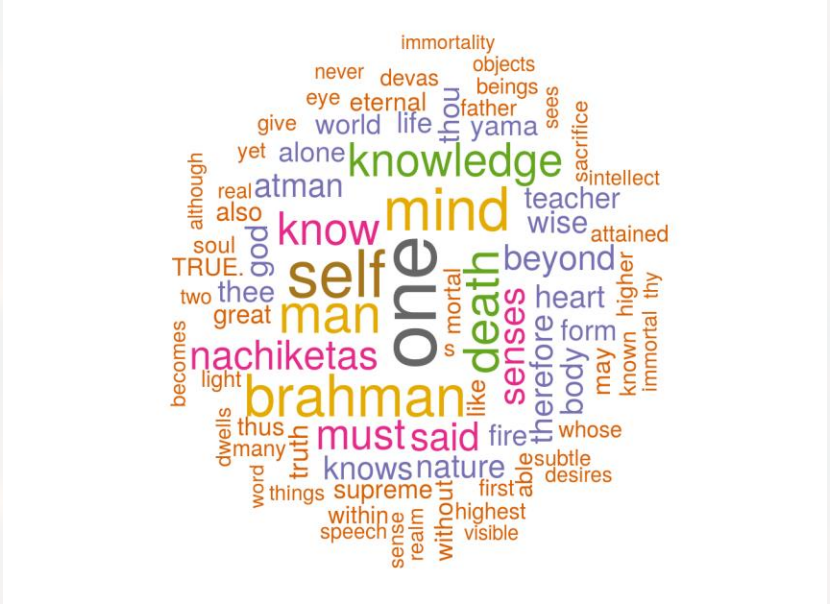
Biblia



Częstość występowania słów - im większe, tym słowo częściej występuje.

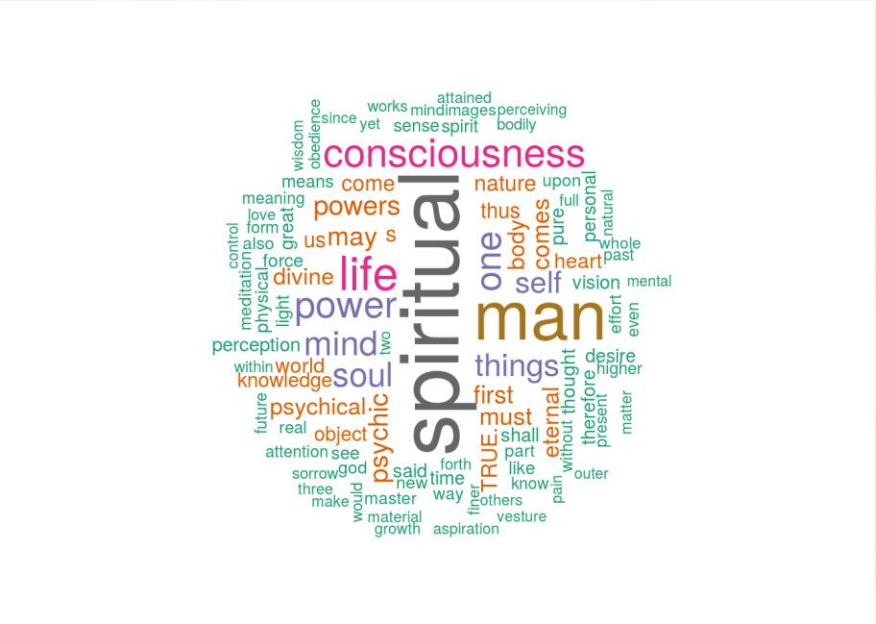
Buddyzm

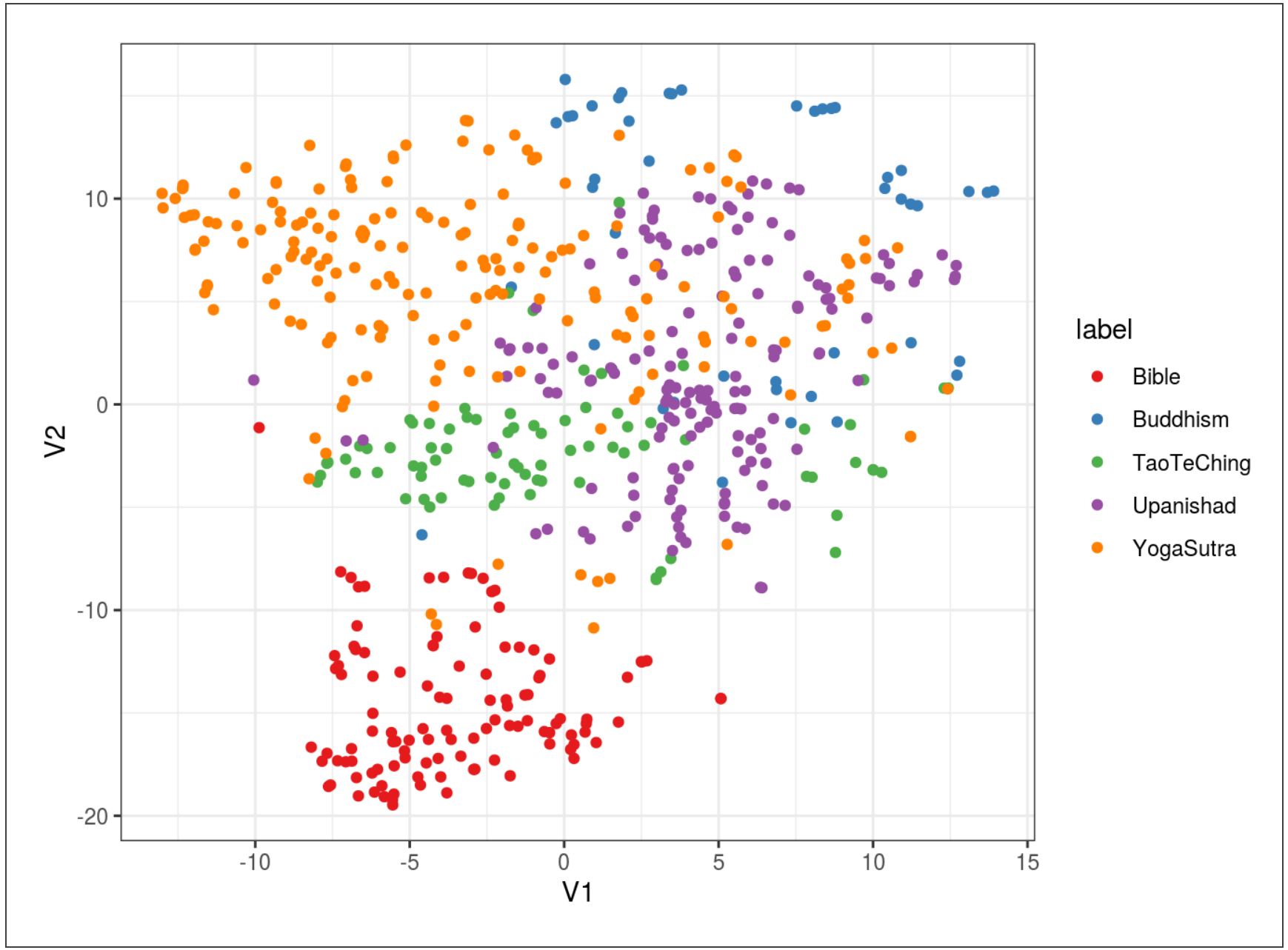
Upniszady



Częstość występowania słów - im większe, tym słowo częściej występuje.

Yoga sutry




10
0
-10
-20
V2
-10
-5
0
5
10
15
V1
label
Bible
Buddhism
TaoTeChing
Upanishad
YogaSutra

TSNE

Bibilę da się wyodrębnić na tle innych ksiąg

Łączenie 4 ksiąg w Biblię pozwoli osiągnąć lepsze wyniki

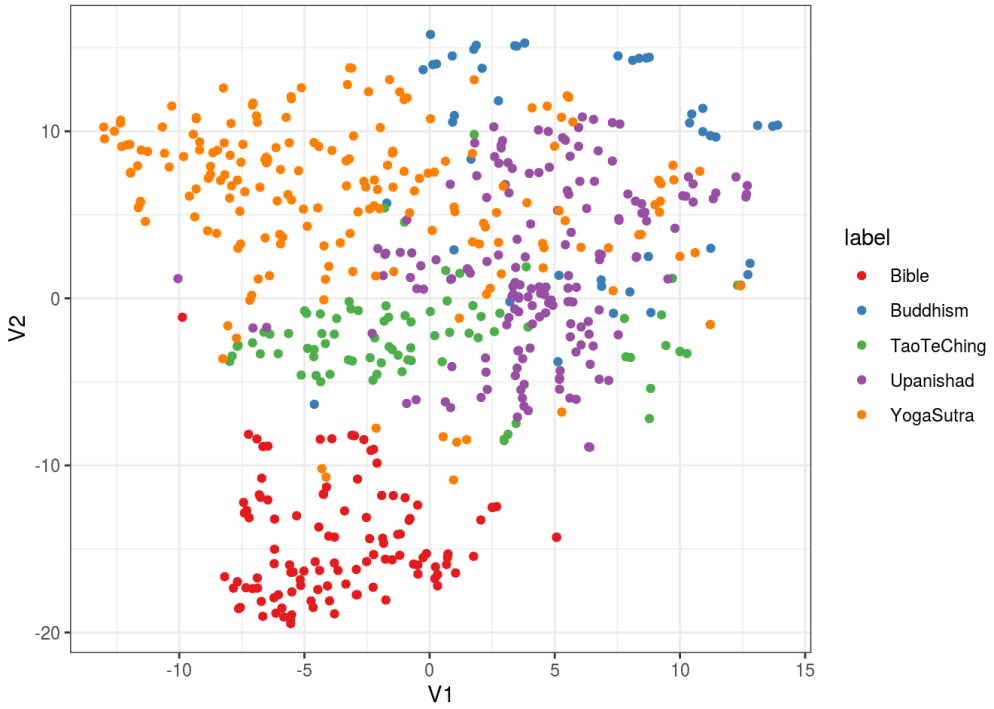
Inżynieria cech

Do inżynierii cech engineering spróbowaliśmy następujących podejść:

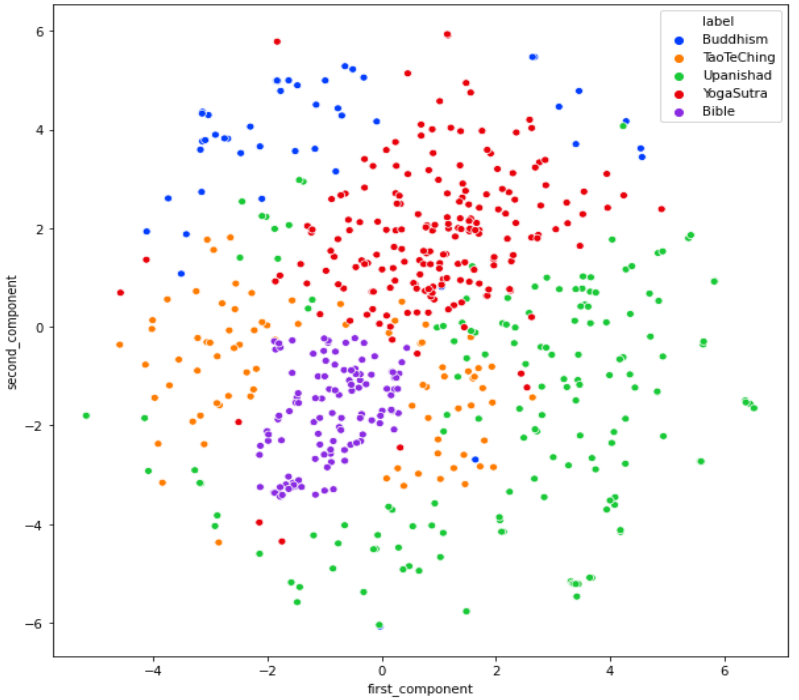
- **stop_words** - usunięcie z kolumn słów pokroju "the", "and" itp.
- entities – encja, nazwa, reprezentująca rzeczywiste obiekty (osoby, organizacje, jednostki geopolityczne etc...)
- **słownika nlp (spacy)** - biblioteka, która zawiera własny słownik, wyodrębniający słowa z tekstu
- **TfidfTransformer** - skaluje wystąpienia biorąc pod uwagę wystąpienia w całym tekście, by uwzględnić wagę i częstotliwość danego słowa
- lematyzacja - sprowadzanie słów do ich podstawowego słowa, np. kotek -> kot, mierzyła -> miara
- stemming - usuwanie końcówek fleksyjnych np. kotek -> kot, mierzyła -> mierz
- usuwanie kolumn o niskiej wariancji
- redukcja wymiarów przy pomocy tsne i pca

użyte przekształcenia

Przed



Po



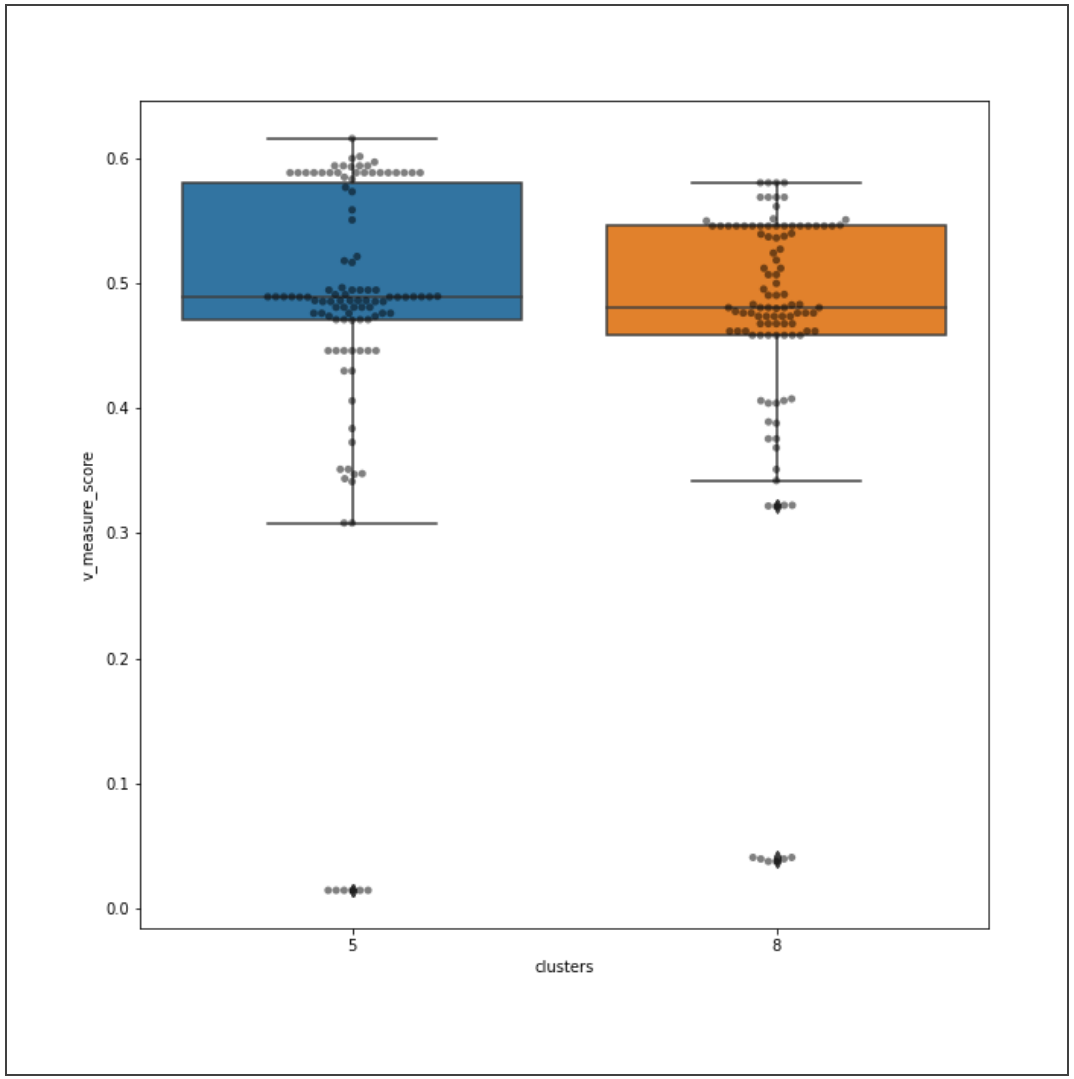
SKUTEK

Użyte Modele

- KMeans
- MiniBatchKMeans
- DBSCAN
- Birch
- AgglomerativeClustering
- SpectralClustering
- GaussianMixture

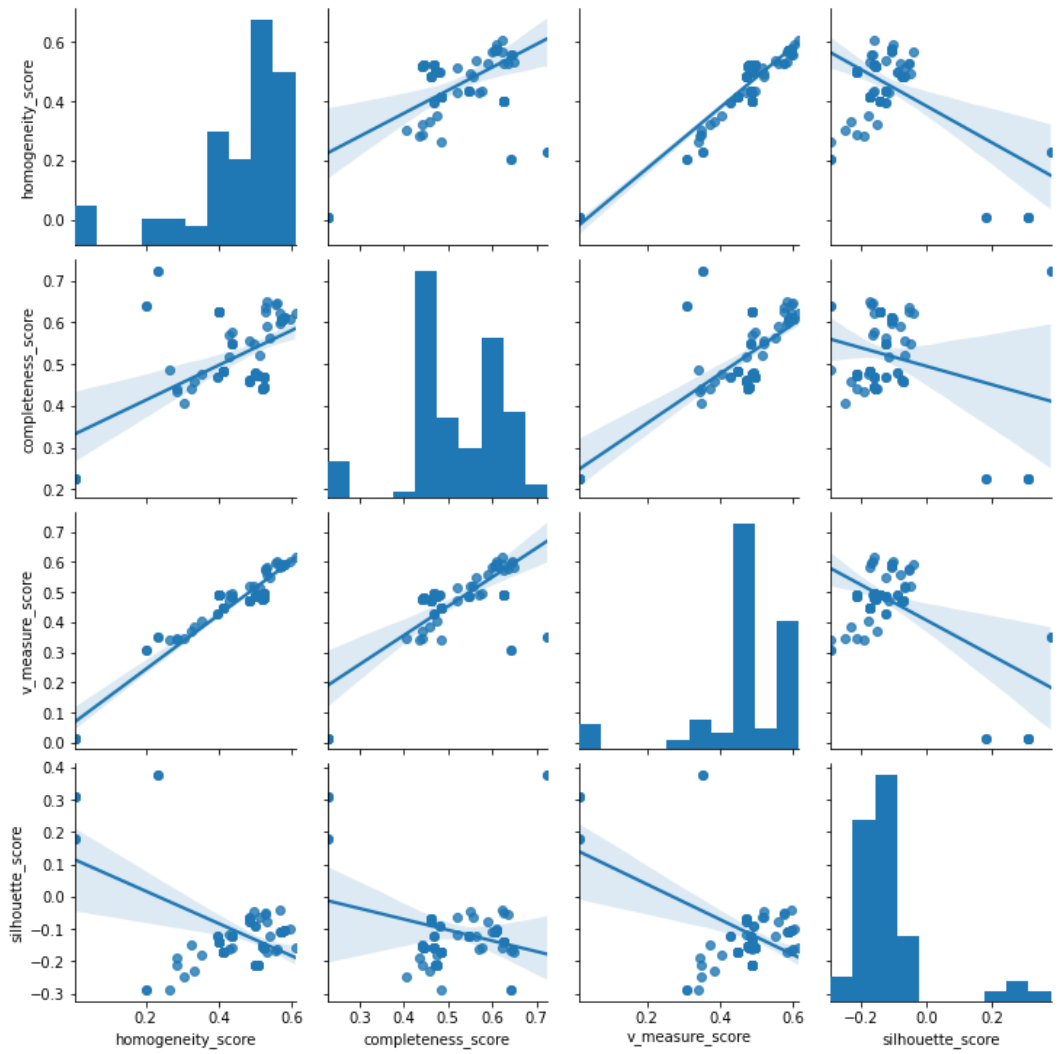
Metryki

- **homogeneity_score** - mówimy, że przyporządkowanie jest homogeniczne, gdy w każdym klastrze są obserwacje tylko jednej klasy. Im bliżej 1.0 tym lepiej.
- **completeness_score** - mówimy że przyporządkowanie jest kompletne, gdy w wszystkie obserwacje danej klasy są przyporządkowane do tego samego klastra. Im bliżej 1.0 tym lepiej.
- **v_measure_score** - średnia harmoniczna z dwóch powyższych.
- **silhouette_score** - ta miara mówi nam jak bardzo obserwacje w jednym klastrze są do siebie podobne i jak bardzo różnią się od obserwacji w innych klastrach. Wynik bliski 1.0 mówi nam, że obserwacje wewnątrz klastrów są do siebie podobne i różnią się od tych z innych klastrów.



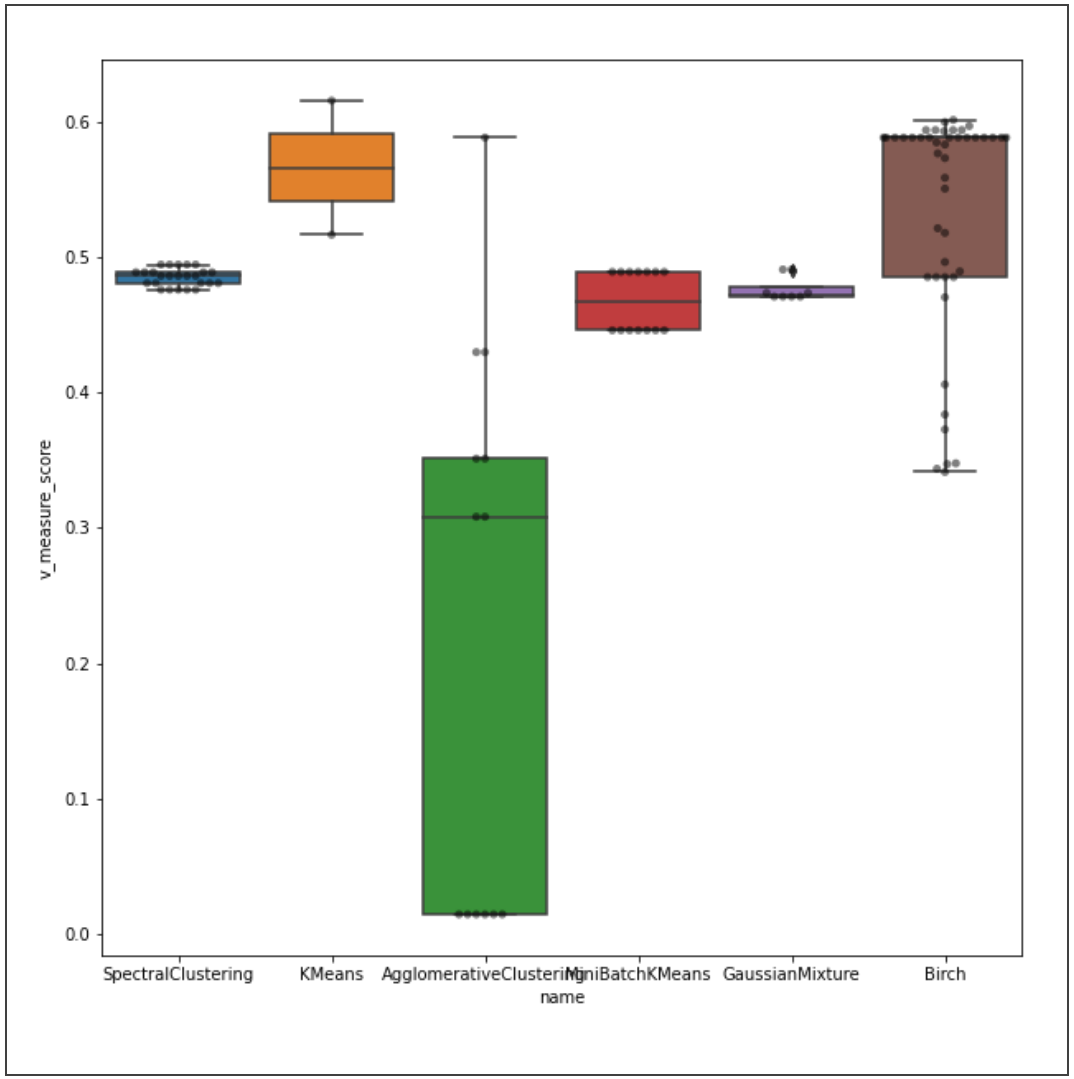
Strojenie hiperparametrów

- Dla każdego modelu wybraliśmy 2 parametry a następnie autorską funkcją grid search sprawdziliśmy ich wszystkie kombinacje
- Eksperyment zrobiliśmy dla 2 różnych liczb klastrów 5 oraz 8.
- Ostatecznie wybraliśmy miarę v_measure_score jako ostateczną. Dlaczego akurat ta miara?



Wnioski ze strojenia hiperparametrów

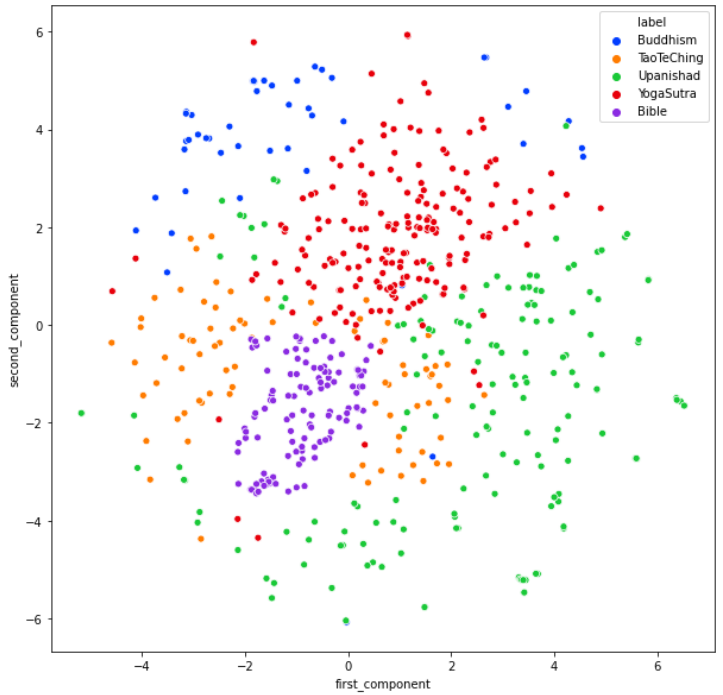
- Najbardziej skorelowana z innymi jest metryka v_measure_score – dlatego będziemy jej używać
- Slihouette score jest dla tych danych niemiarodajny, klastry są ze sobą pomieszane
- Najlepszym modelem jest…



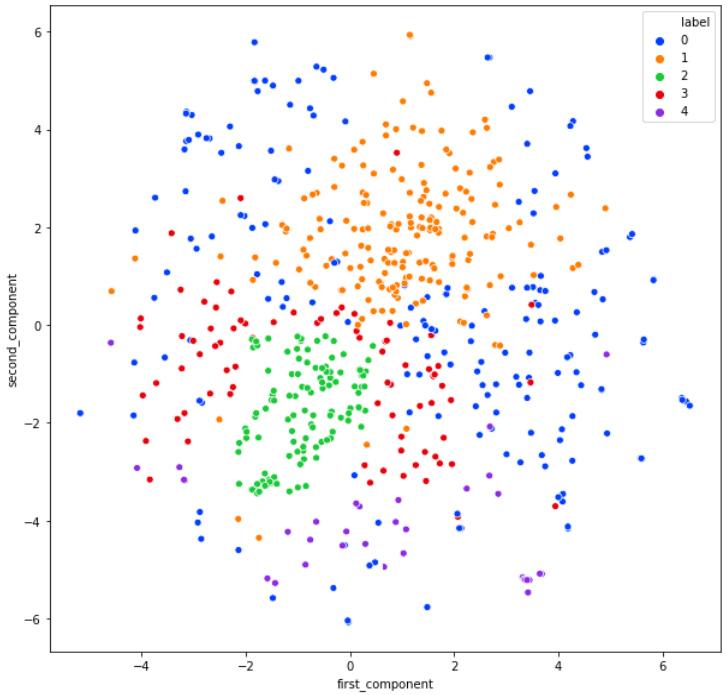
Najlepsze metody

- Najlepszy wynik miał KMeans, ale to dla wyniki **Birch**'a były na tyle dobre, stabilne, i niepodatne na losowość, że to właśnie on zasługuje na miano najlepszego modelu.
- Porównajmy jak klastruje dane po przekształceniu TSNE i jak wyglądają "prawdziwe klastry"

Prawdziwe



Birch



Oznaczenie Biblii

100% poprawne

… to czy Biblię można rozpoznać zawsze?

- Można, I to jak! Birch wyodrębnił Biblię bezbłędnie.

Q & A


Dziękujemy za Waszą uwagę!