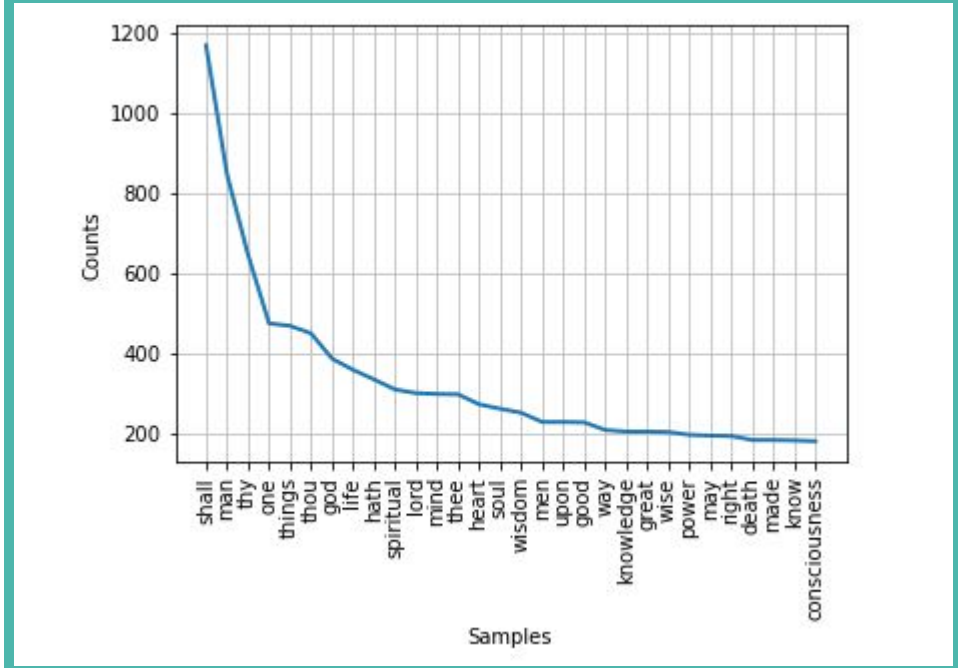
Święte księgi

Staroń, Szypuła, Urbala

Wstępne przekształcenia

- Usunięcie stop words
- Stemming
- TF-IDF

Wstępne analizy



Hipoteza #1

Czy rozdziały klastrują między poszczególne księgi?

Hipoteza #1

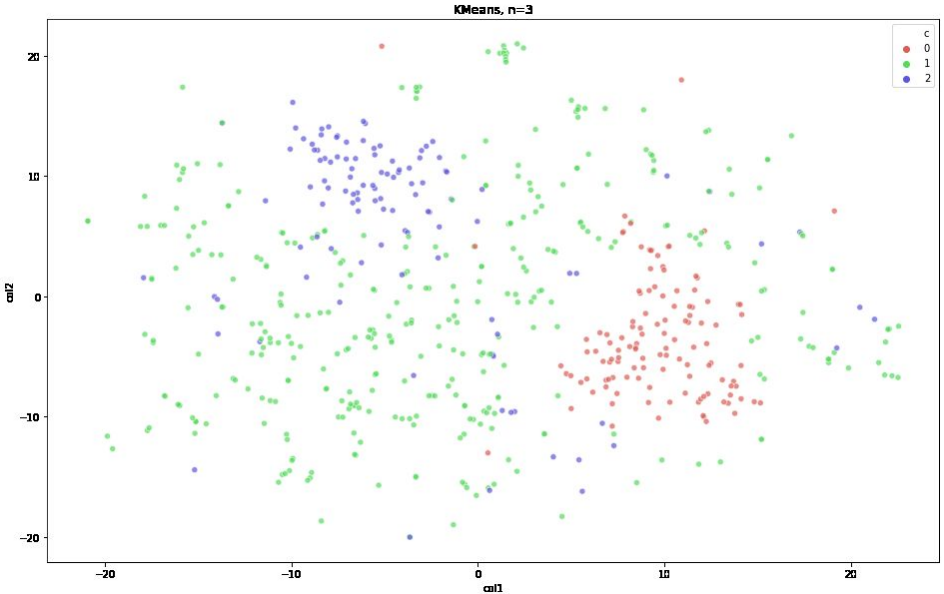
Czy rozdziały klastrują między poszczególne księgi?

Niestety nie.

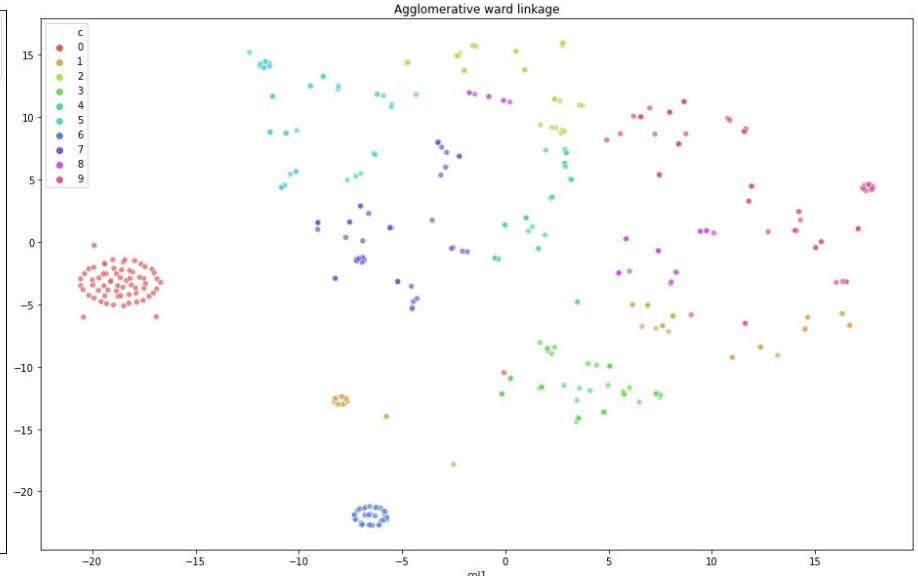
	adjusted_rand_score
Agglomerative ward linkage	0.031006
Agglomerative complete linkage	-0.009290
Agglomerative average linkage	-0.015803
Agglomerative single linkage	-0.003374
Mini Batch KMeans	0.009919
Kmeans	0.019211

Jak liczyć odległości

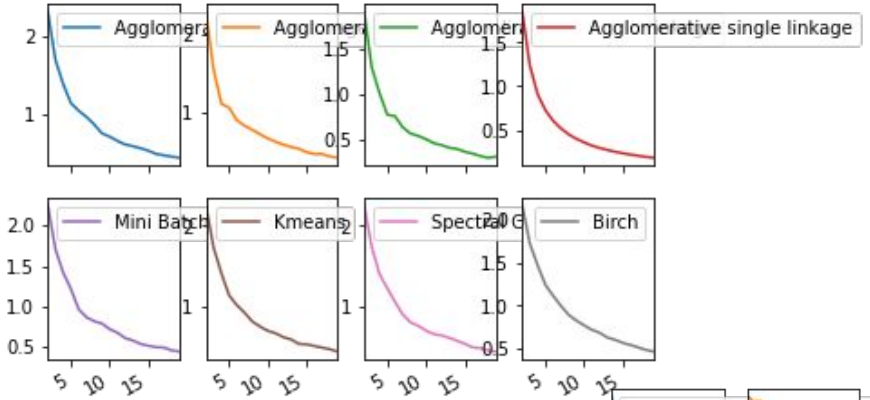
TF-IDF na całym tekście



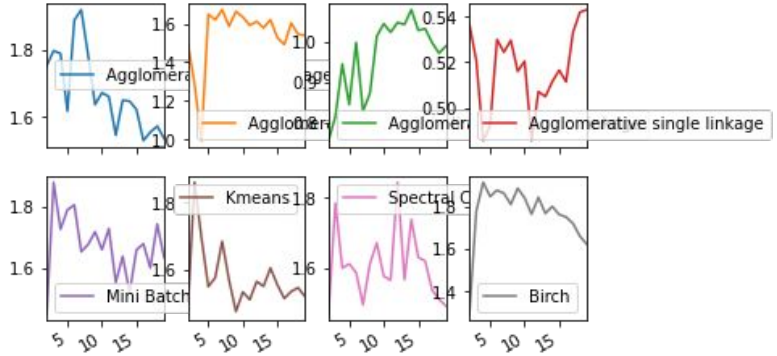
Ograniczenie korpusu do top 30 słów



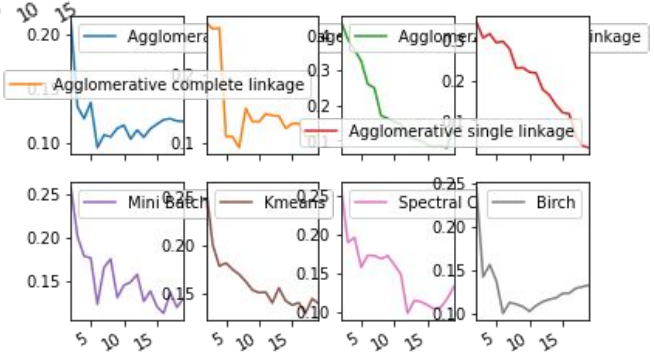
Sentymenty - analiza miar



WCSS

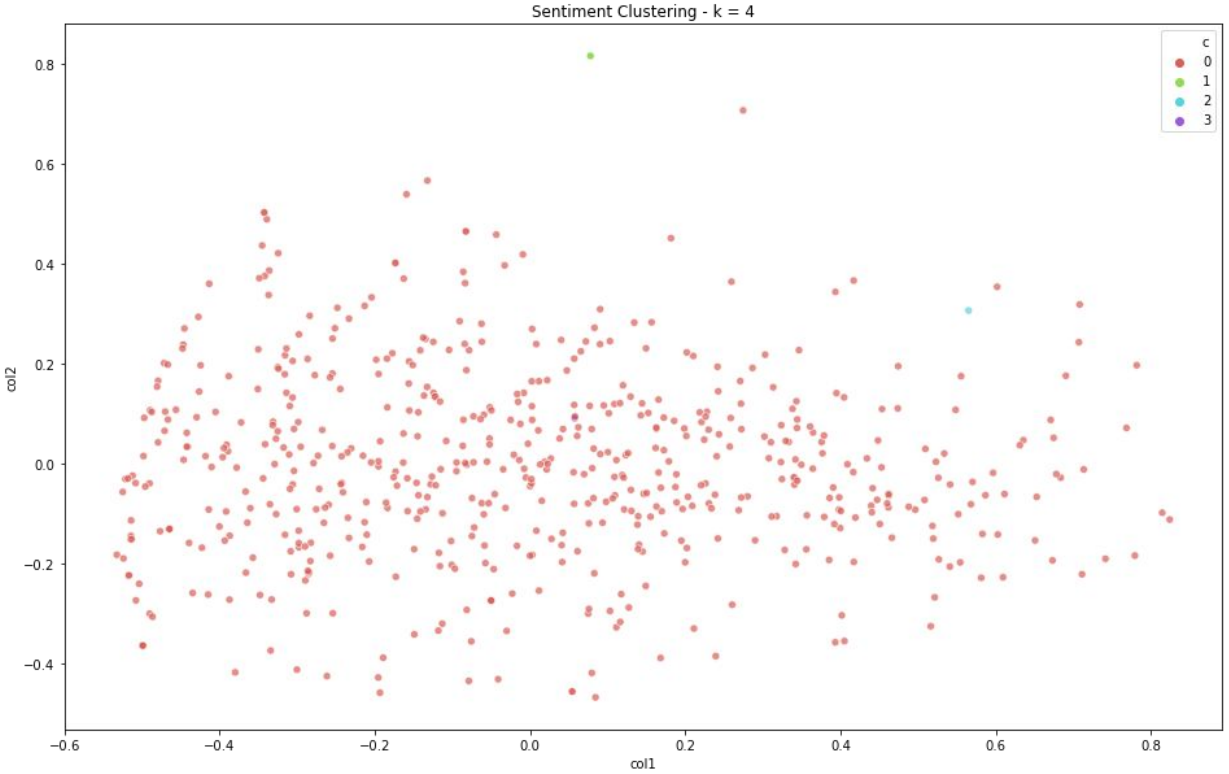


I D-B

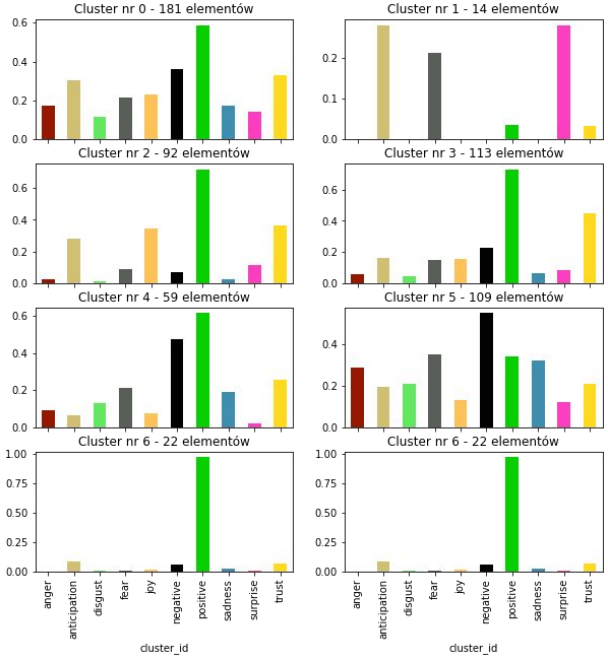


Silhouette

Sentymenty plot - PCA



Sentymenty plot - nasz sposób



Hipoteza #2 - początki i końce

Czy pierwsze i ostatnie rozdziały są powiązane ze sobą?

Hipoteza #2 - początki i końce

Czy pierwsze i ostatnie rozdziały są powiązane ze sobą?

Również, nie.

	adjusted_rand_score
Agglomerative ward linkage	-0.066667
Agglomerative complete linkage	-0.053512
Agglomerative average linkage	0.000000
Agglomerative single linkage	0.000000
Mini Batch KMeans	-0.066667
Kmeans	-0.053512
Spectral Clustering	-0.053512
Birch	-0.053512

Podsumowanie

Analiza i klastrowanie sentymentów nie przyniosły satysfakcjonujących wyników.

Dużo lepsze efekty dało zastosowanie algorytmów klastrujących na zbiorze informacji o częstości występowania słów w rozdziałach.