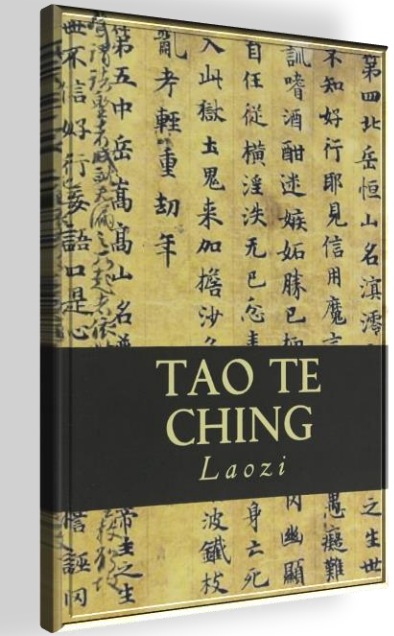
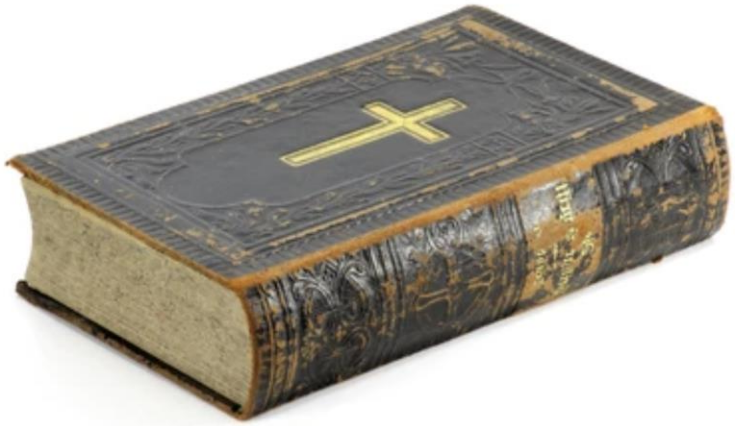
Unsupervised Learning

NLP and Clustering of religious texts



Autorzy:
Adrianna Grudzień
Mateusz Stączek

2021



Dane

- Upanishads,
- Yoga Sutras,
- Buddha Sutras,
- Tao Te Ching,
- Book of Wisdom,
- Book of Proverbs,
- Book of Ecclesiastes,
- Book of Ecclesiasticus.

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

Google

View ALL Data Sets

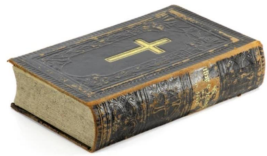
A study of Asian Religious and Biblical Texts Data Set

Download: Data Folder, Data Set Description

Abstract: Mainly from Project Gutenberg, we combine Upanishads, Yoga Sutras, Buddha Sutras, Tao Te Ching and Book of Wisdom, Book of Proverbs, Book of Ecclesiastes and Book of Ecclesiasticus

Data Set Characteristics:	Multivariate, Text	**Number of Instances:**	590	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	8265	**Date Donated**	2019-12-24
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	30993

https://archive.ics.uci.edu/ml/datasets/A+study+of++Asian+Religious+and+Biblical+Texts



Wgląd w dostępne dane

- Surowy tekst

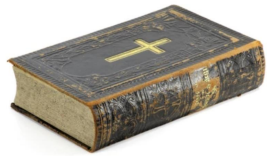
- Bag of Words (8266 words)

```
precious stones.  Verses XVI-XVIII are regarded by many as an
interpolation, which would account for certain obscurities and
repetitions in them.
2.37
Nachiketas said: There is this doubt regarding what becomes of
a man after death.  Some say he exists, others that he does
not exist.  This knowledge I desire, being instructed by
thee.  Of the boons this is the third boon.
2.38
Yama replied: Even the Devas (Bright Ones) of old doubted
regarding this.  It is not easy to know; subtle indeed is this
subject.  O Nachiketas, choose another boon.  Do not press me.
Ask not this boon of me.
2.39
Nachiketas said: O Death, thou sayest that even the Devas had
doubts about this, and that it is not easy to know.  Another
teacher like unto thee is not to be found.  Therefore no other
boon can be equal to this one.
2.40
Yama said: Ask for sons and grandsons who shall live a hundred
years, many cattle, elephants, gold and horses.  Ask for lands
of vast extent and live thyself as many autumns as thou
desirest
```

```
df_csv = pd.read_csv('AllBooks_baseline_DTM_Unlabelled.csv')
df_csv.head(2)
```

	# foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	convict	...	erred	thin
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

2 rows × 8266 columns



Nasze dane

- Dane statystyczne
- Bag of Words (4285 words)
- Robust Scaler
- PCA

Preprocessing

Next, we prepare a data frame `df_chapters` containing

- chapter length,
- number of sentences,
- average sentence length,
- number of words (without stopwords),
- number of unique words (without stopwords),
- average word length,
- chapter complexity (flesch reading ease),
- positive/negative/neutral tinting of chapter,
- bag of words instead of texts.

Nazwy kolumn

```
['num_of_sentences',
 'avg_sentence_len',
 'num_of_words',
 'num_of_words_wo_stopwords',
 'num_of_uniq_words',
 'num_of_uniq_words_wo_stopwords',
 'num_of_letters',
 'avg_word_length',
 'text_complex_fre',
 'polarity',
 'subjectivity',
 'aaron',
 'abandoned',
 '...',
 'yoga',
 'yoke',
 'young',
 'youth',
 'zeal']
```

Pierwsze kilka kolumn poniżej:

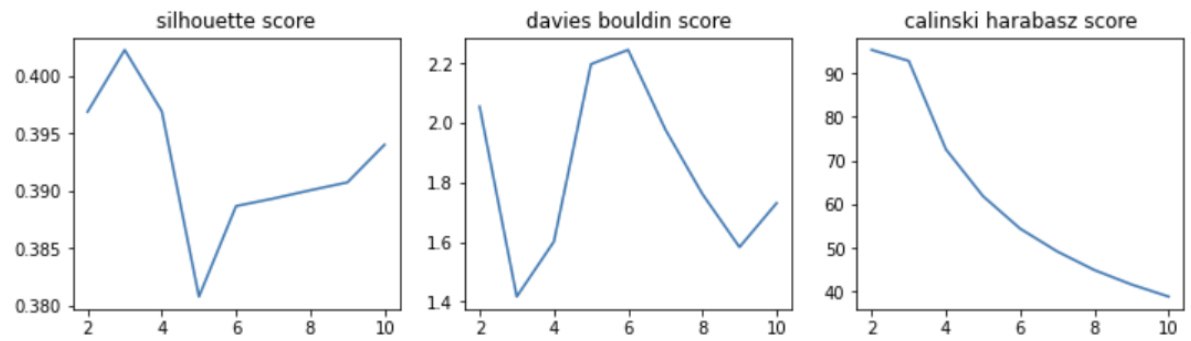
	chapter_length	num_of_sentences	avg_sentence_len	num_of_words	num_of_words_wo_stopwords	num_of_uniq_words	num_of_uniq_words_wo_stopwords
0	3628	28	129.035714	602	299	168	85
1	1509	15	99.933333	265	107	101	48

Modele – szukanie liczby klastrów

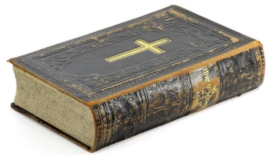
- KMeans – 2 lub 3

- Agglomerative Clustering – 3?

```
plot_metrics(df_csv, AgglomerativeClustering)
```

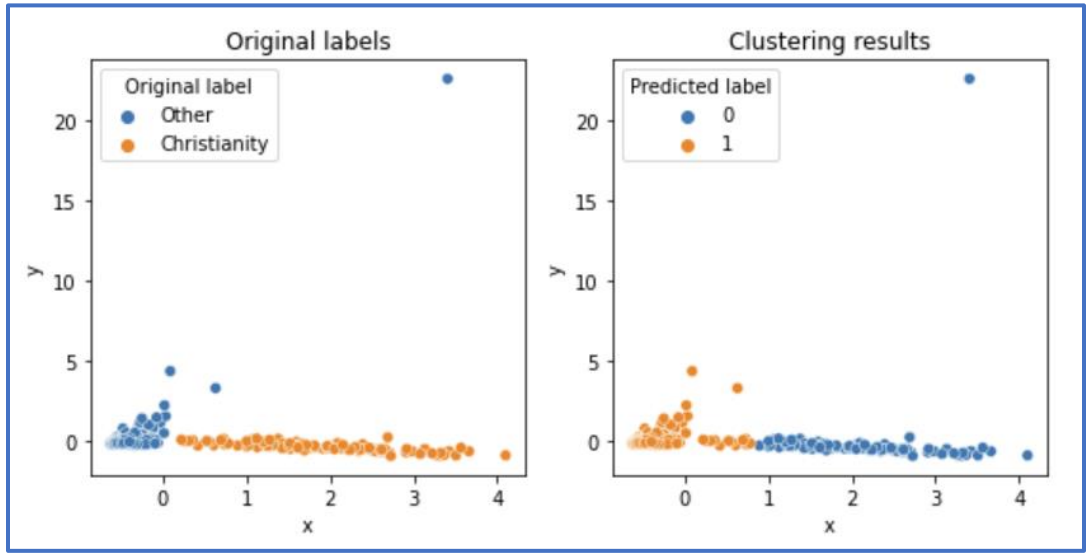


- Gaussian Mixture Model - ?

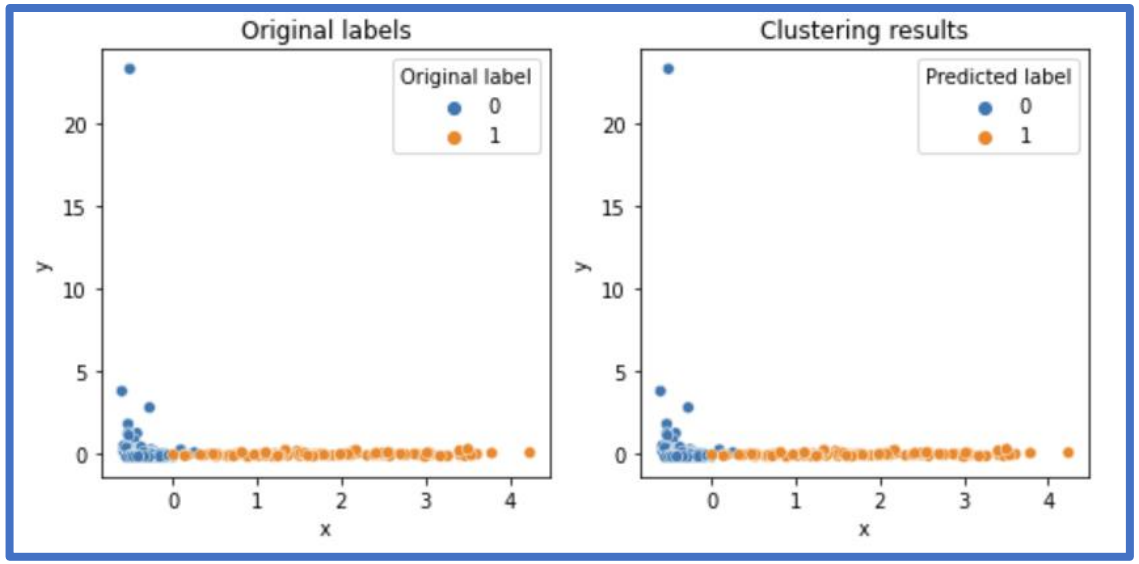


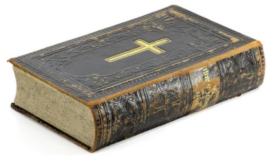
Modele - wizualizacja

KMeans



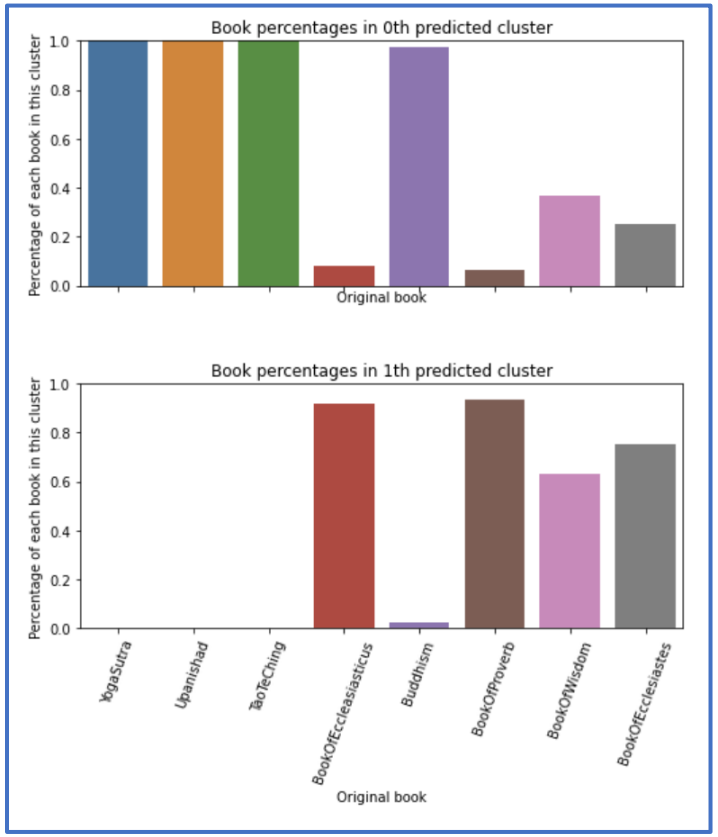
Agglomerative Clustering

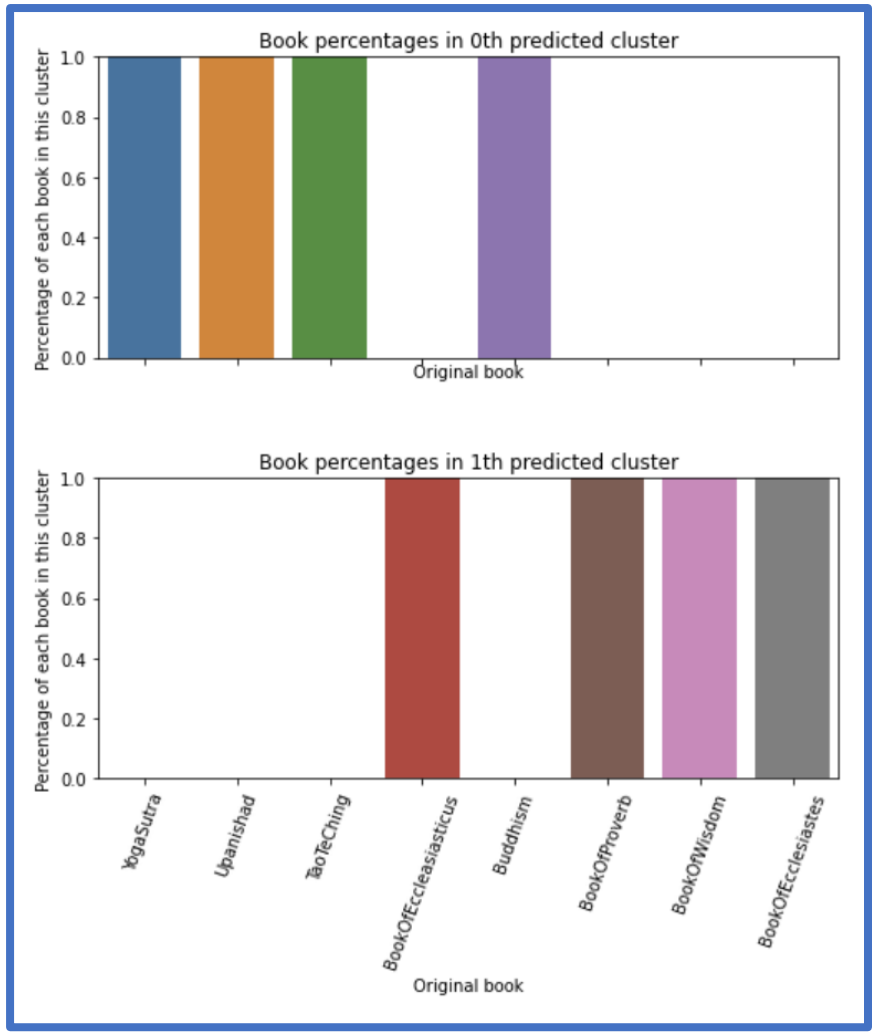




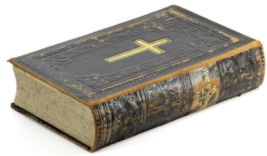
Modele – wizualizacja

KMeans





Agglomerative Clustering



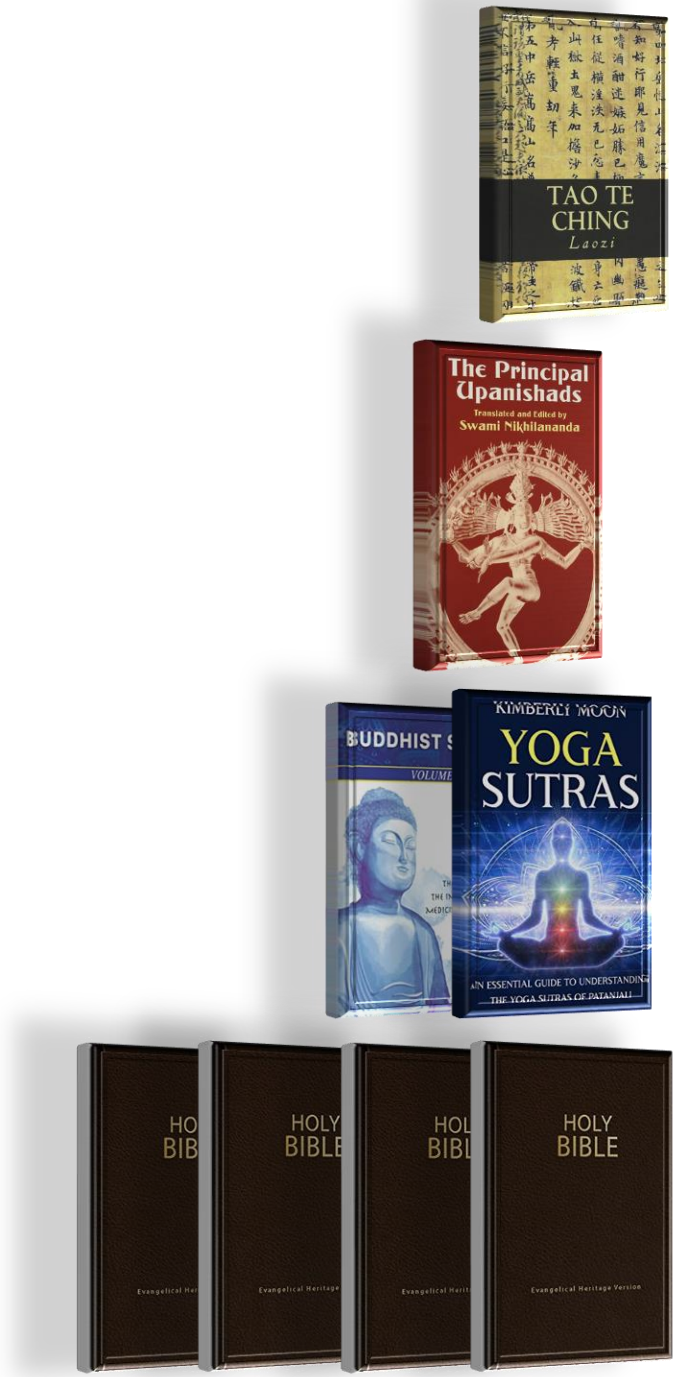
Podsumowanie i dalsze badania

Działa:

- Podział na 2 klastry – chrześcijańskie vs wschodnie

Co warto sprawdzić:

- Przetwarzanie surowego tekstu
- Skalowania różne
- Modele bez podania liczby klastrów



Dziękujemy za uwagę

Skąd powyższy tekst? :P