EKSPLORACYJNA ANALIZA DANYCH

PROJEKT 2, KAMIEŃ MILOWY 1

DOMINIK PAWLAK, PRZEMYSŁAW OLENDER

WPROWADZENIE

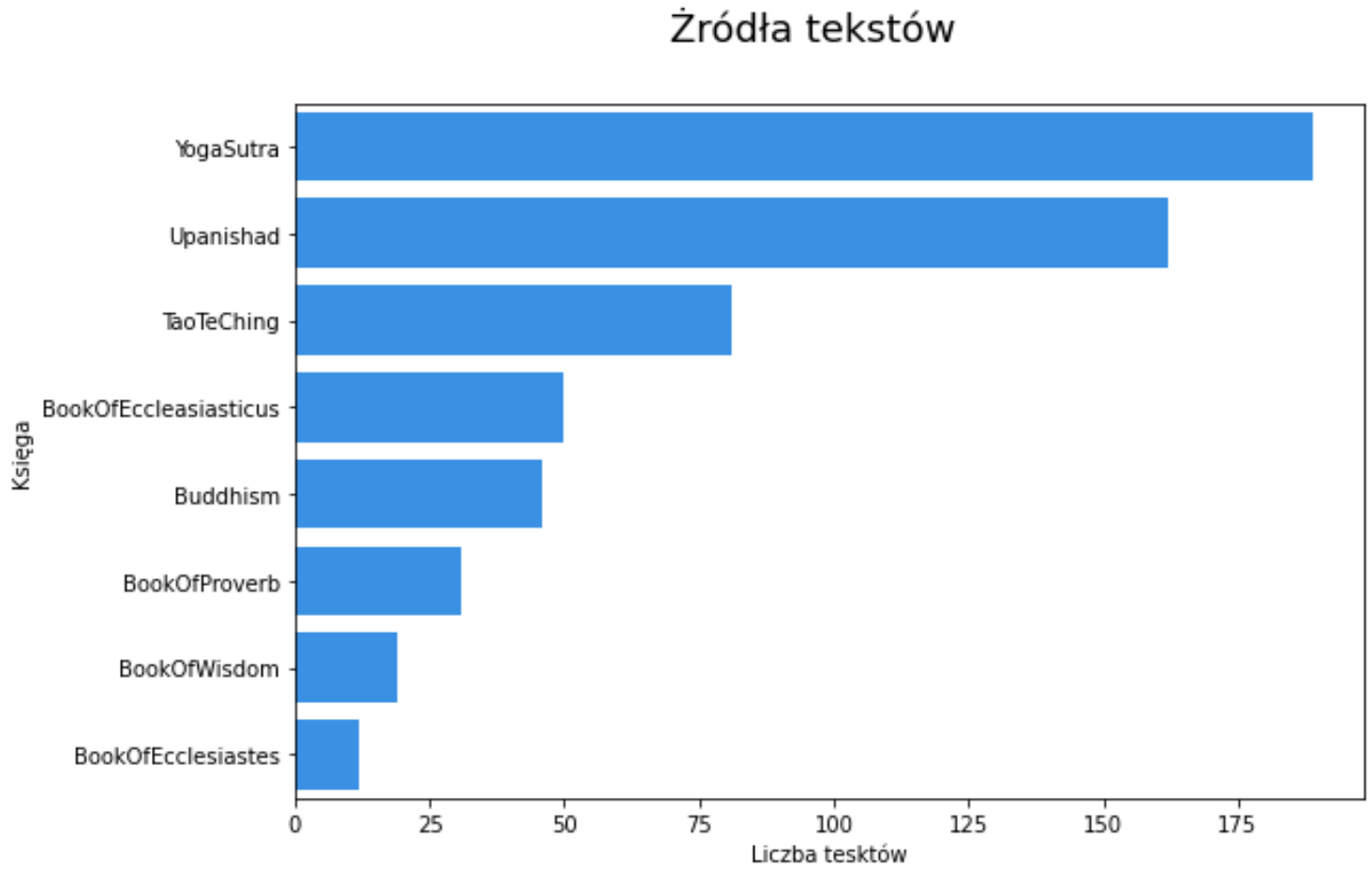
- Tematem naszej pracy jest klasteryzacja tekstów pochodzących z 8 Świętych ksiąg 4 różnych religii:
  - Chrześcijaństwo - Księgi ze Starego Testamentu: Book of Proverbs, Book of Wisdom, Book of Ecclesistes, Book of Eccleasiasticus
  - Hinduizm: Yoga Sutra, Upanishad
  - Buddyzm – jedna księga
  - Taoizm – jedna księga
- Zbiór danych zawiera 3 pliki:
  - Plik .txt z pełnymi tekstami
  - Plik .csv z tabela słów występujących w danych tekstach
  - Plik .csv z tabelą słów występujących w tekstach i informacją o pochodzenie tekstu

DANE

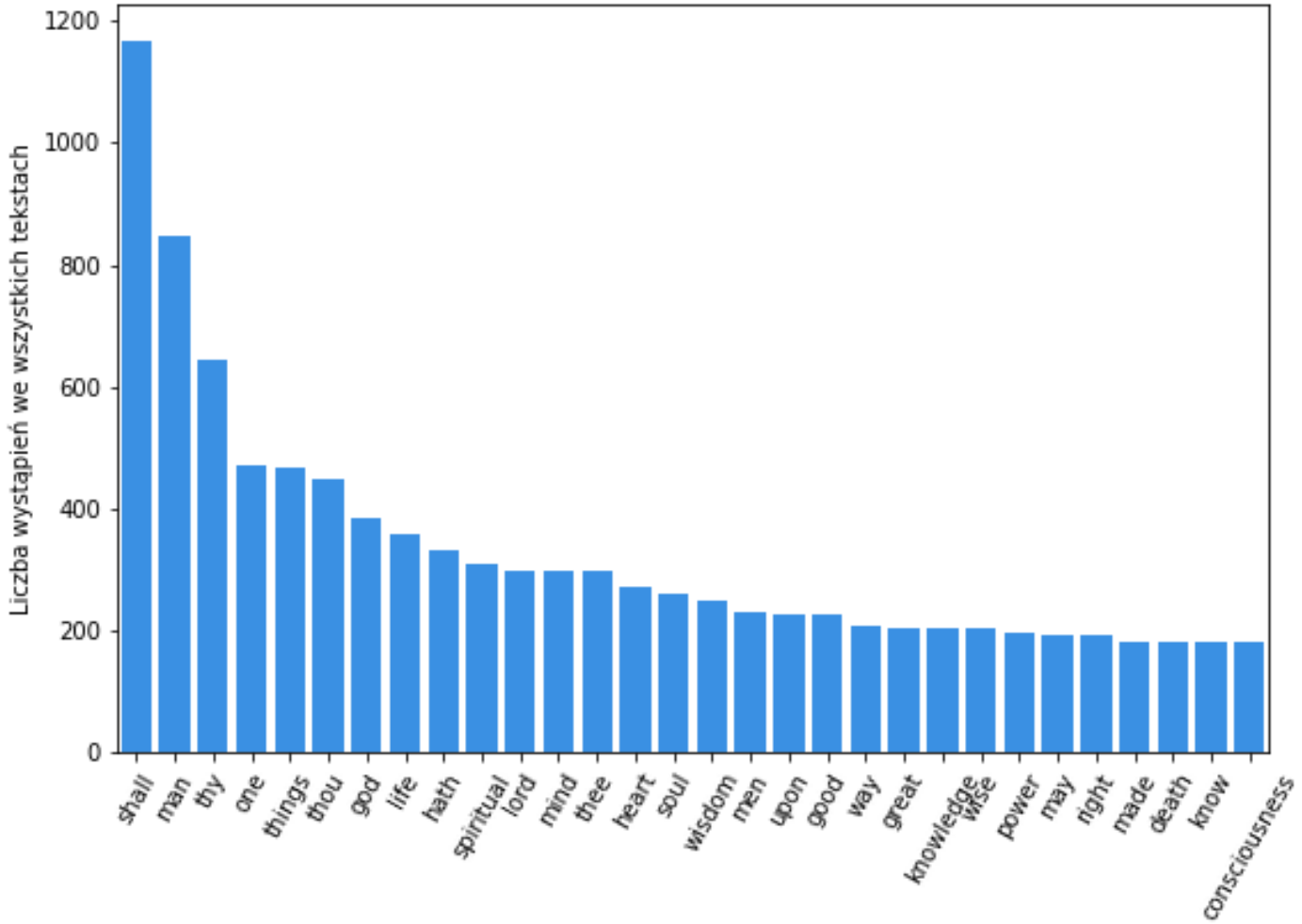
- Ramka danych z podpisanymi kolumnami zawiera:
  - 590 wierszy – tyle ile tekstów
  - 8267 kolumn – tyle ile unikalnych słów znajduję się w tekstach + jedna kolumna z informacją o tekście.
  - W komórce wspólnej dla tekstu i słowa mamy liczbę jego wystąpień

	Unnamed: 0	foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	...	erred	thinkest	modern	reigned	sparingly	visual
0	Buddhism_Ch1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
1	Buddhism_Ch2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	Buddhism_Ch3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	Buddhism_Ch4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	Buddhism_Ch5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

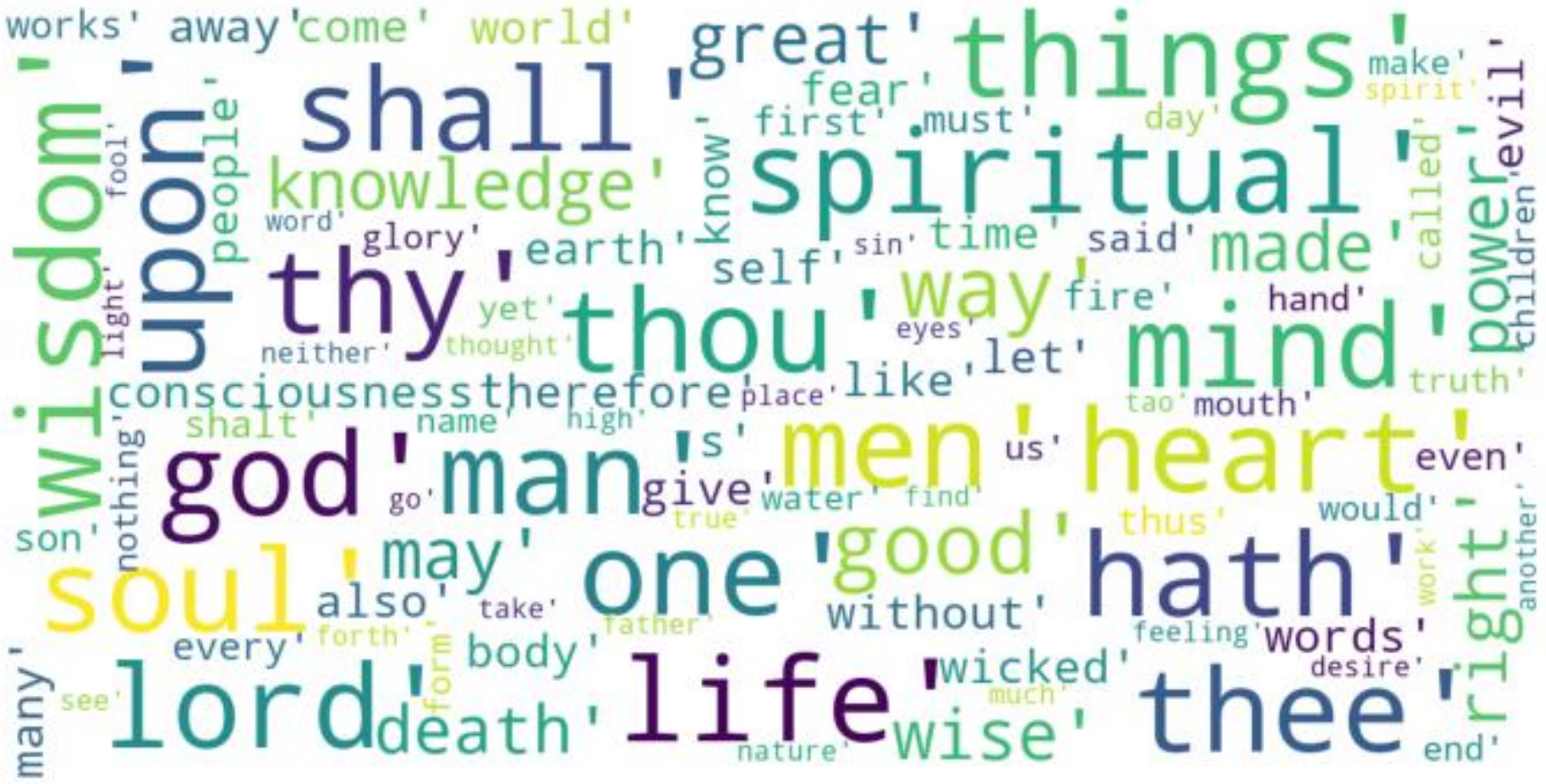
LICZBA WYSTĄPIEŃ TEKSTÓW Z DANEJ RELIGII



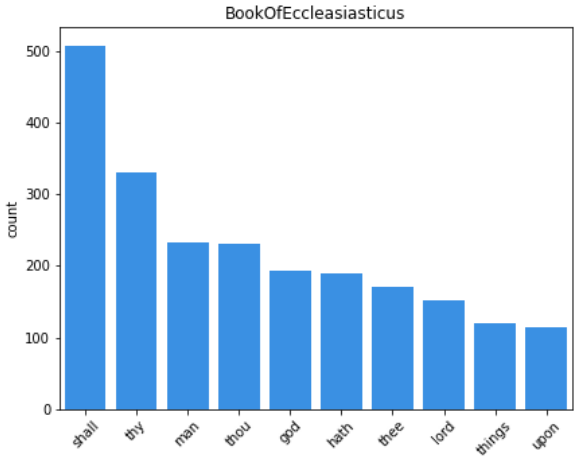
NAJPOPULARNIEJSZE SŁOWA

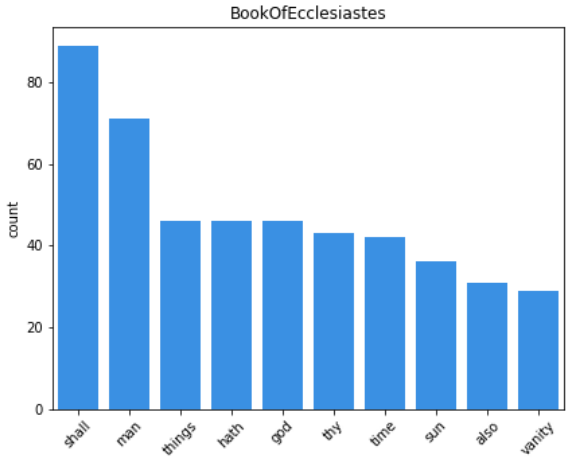


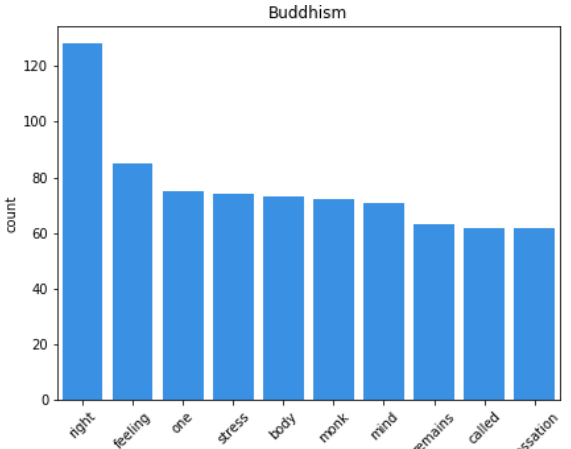
NAJPOPULARNIEJSZE SŁOWA

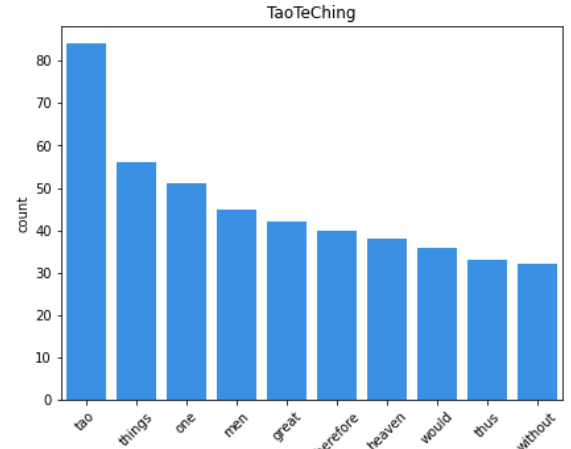


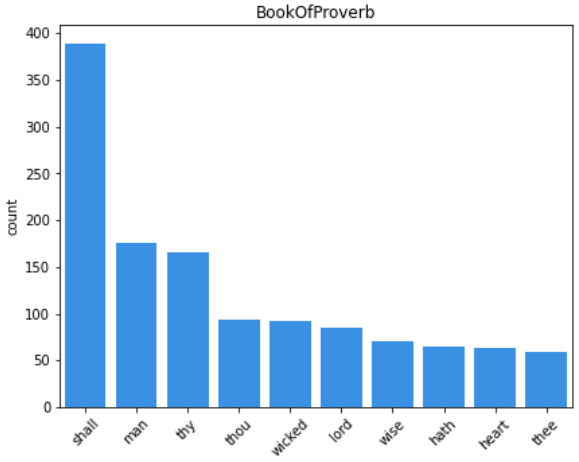
NAJPOPULARNIEJSZE SŁOWA DLA DANEJ RELIGII

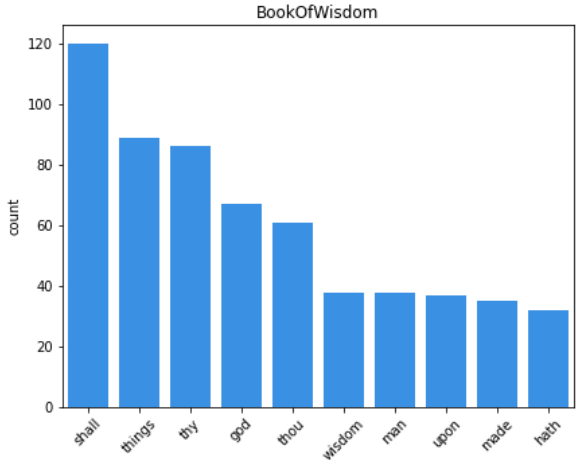


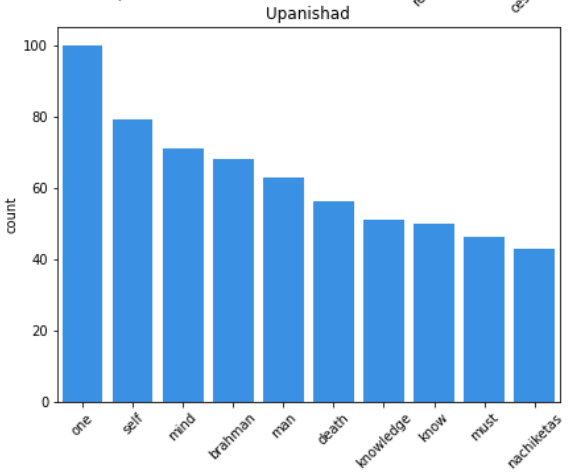


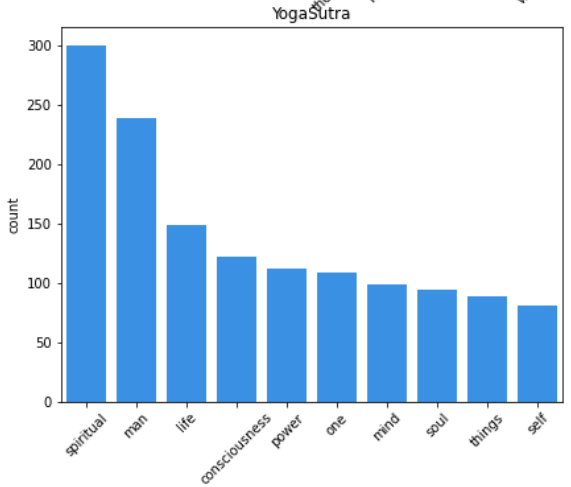




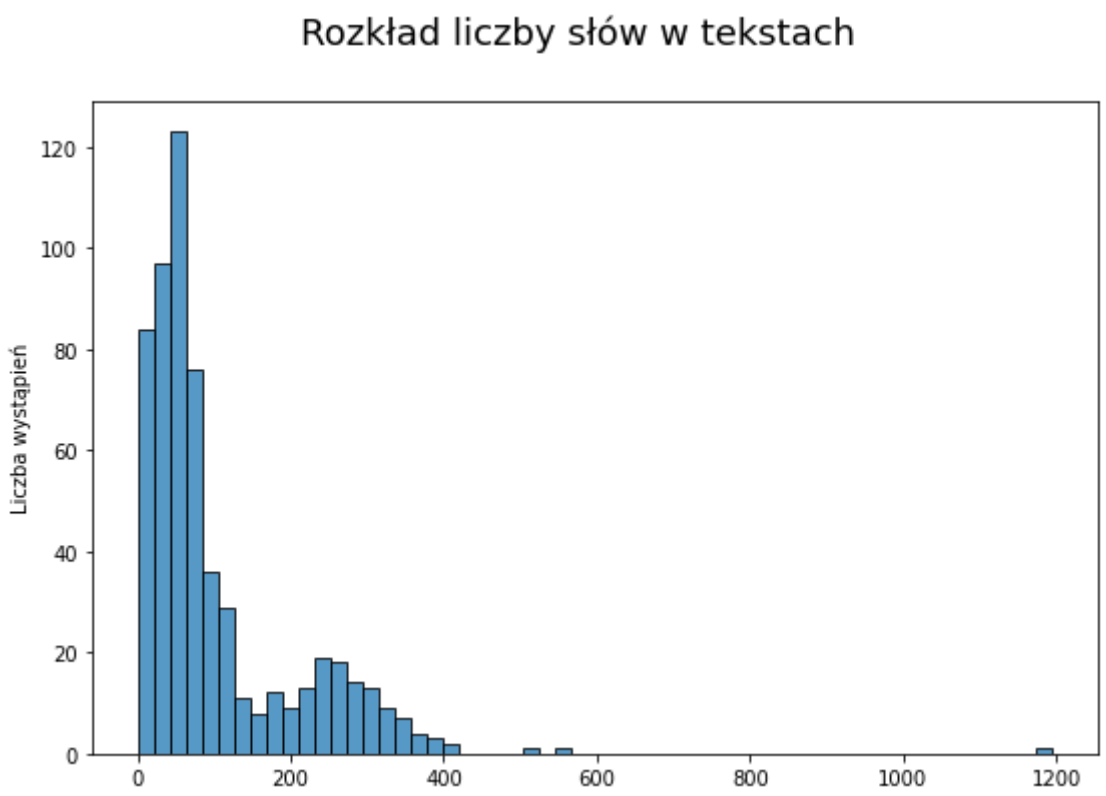


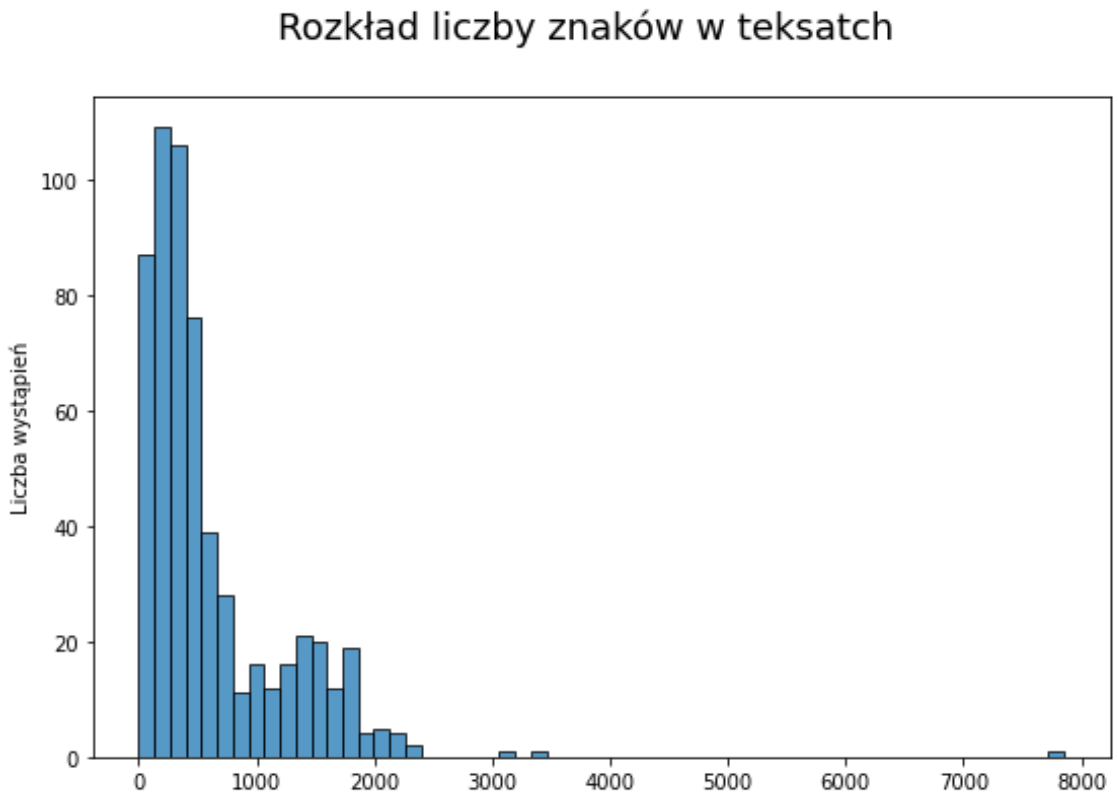






ROZKŁAD DŁUGOŚCI TEKSTÓW





Dane bez etykiet

```
1 print(f"Kształt ramki: {df.shape}.")
2 df.head()
```

```
Kształt ramki: (590, 8266).
```

	foolishness	hath	wholesome	takest	feelings	anger	vaivaswata	matrix	kindled	convict	...	erred	thinkest	modern	reigned	sparingly	visual	though
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	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	0.0	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0

Mamy dane bez etykiet. 590 rekordów (źródeł danych) oraz 8266 kolumn. Nie mamy braków. Wszystkie dane są liczbami naturalnymi, głównie są to 0.

Wstępne EDA

- W pierwszej kolejności sprawdziliśmy czy występują skrótowce, tj. słówka typu 'don't', 'isn't', 'aren't' itp. Okazało się, że takich słówek nie ma.
- Następnie postanowiliśmy zmienić nasze słowa, tak żeby mieć tylko ich podstawę słowotwórczą. Tj. zamienić "playing" -> "play" itp. Wykorzystaliśmy do tego bibliotekę Spacy. Stworzyliśmy słownik, w którym klucze to słówko oryginalne, nt. Wartość to jego podstawa słowotwórcza.

Słowotwórstwo w analizie danych

- Okazało się to dobrym pomysłem, ponieważ otrzymany słownik miał ponad 2700 kluczy.
- Pozwoliło to znacznie zmniejszyć rozmiar naszej ramki danych, ponieważ niektóre słowa istniały w ramce więcej niż jeden raz, były zapisane w innej formie.

```
Klucz: opposed, Wartość: oppose.
Klucz: opposing, Wartość: oppose.
```

- Po zmienieniu nazw kolumn natrafiliśmy na problem, że mieliśmy
- zduplikowane nazwy kolumn.

```
Kolumny unikalne: 6277.
Wszystkie kolumny: 8266.
```

- Zsumowanie danych ze zduplikowanych kolumn znacznie zmniejszyło
- rozmiar ramki.

```
Nowy kształt ramki: (590, 6277)
```

```
1 df["oppose"]
```

	oppose	oppose
0	0.0	0.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
...	...	...
585	0.0	0.0
586	0.0	0.0
587	0.0	0.0
588	0.0	0.0
589	0.0	0.0

590 rows × 2 columns

Usunięcie stopwords

- Stopwords, to słowa najpopularniejsze w danym języku, nie wnoszące żadnej wartości merytorycznej. W języku angielskim są to słowa typu: “the”, “is”, “in”, “for”, “where”, “when”, “to”, “at”. Sprawdziliśmy czy posiadamy je w naszej ramce.

```
['neither', 'something', 'I', 'elsewhere', 'thus', 'give', 'although', 'perhaps', 'well', 'take', 'out', 'keep', 'thence', 'ful
l', 'nowhere', 'name', 'doing', 'move', 'part', 'nine', 'become', 'for', 'side', 'much', 'someone', 'whole', 'show', 'sometime
s', 'third', 'still', 'empty', 'say', 'see', 'we', 'either', 'will', 'twelve', 'two', 'hereafter', 'might', 'whither', 'who',
'go', 'seem', 'mine', 'bottom', 'beyond', 'as', 'whatever', 'next', 'do', 'down', 'please', 'never', 'therefore', 'get', 'put',
'upon', 'amount', 'formerly', 'within', 'always', 'could', 'front', 'former', 'though', 'towards', 'once', 'often', 'whose', 'a
long', 'already', 'make', 'amongst', 'there', 'call', 'whereas', 'whether', 'behind', 'moreover', 'afterwards', 'in', 'anywher
e', 'all', 'without', 'however', 'may', 'back', 'enough', 'many', 'five', 'ten', 'anyone', 'ever', 'he', 'last', 'other', 'besi
des', 'eleven', 'least', 'also', 'throughout', 'less', 'another', 'toward', 'everywhere', 'must', 'anything', 'quite', 'besid
e', 'hereby', 'almost', 'six', 'thereby', 'nothing', 'alone', 'rather', 'becoming', 'everything', 'top', 'wherever', 'whoever',
'first', 'together', 'wherein', 'this', 'due', 'among', 'namely', 'yet', 'nevertheless', 'beforehand', 'none', 'latter', 'thre
e', 'would', 'eight', 'except', 'several', 'around', 'thereafter', 'even', 'unless', 'the', 'at', 'else', 'one', 'really', 'bei
ng', 'whereby', 'sometime', 'therein', 'hence', 'hundred', 'four', 'such', 'various', 'per', 'just', 'indeed', 'whence', 'other
wise', 'whenever', 'every', 'since', 'everyone']
```

Stopwords

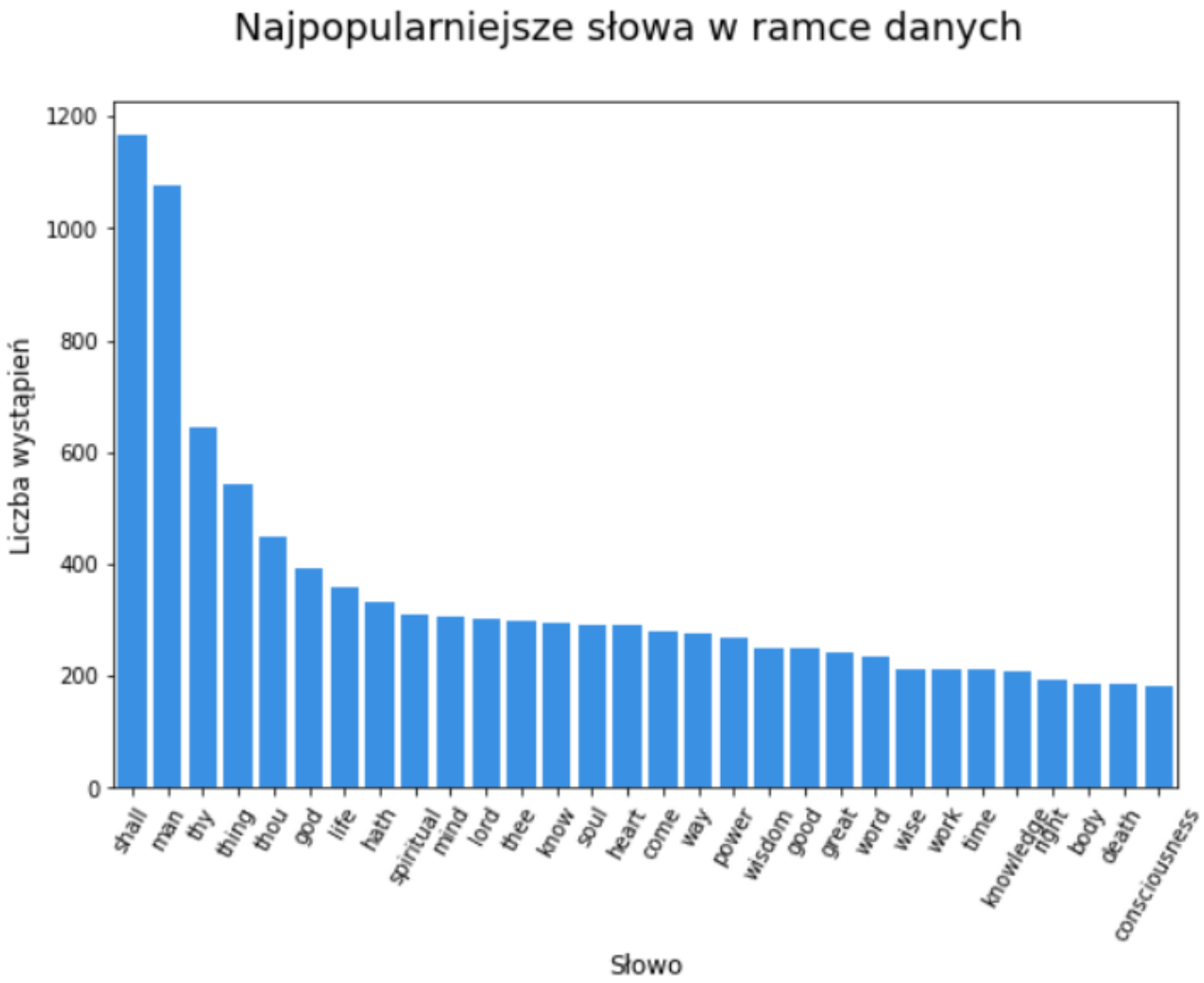
- Otrzymaliśmy, że w naszej ramce znajduje się 166 słów z tej kategorii. Spośród 6000 posiadanych słów, stanowią one niewielki odsetek, więc zdecydowaliśmy się je usunąć.
- Pozwoliło nam to jeszcze bardziej zmniejszyć rozmiar ramki danych.

```
Nowy rozmiar ramki: (590, 6111).
```

- Następnie zbadaliśmy liczbę wystąpień jednego słowa.

```
Średnia liczba wystąpień jednego słowa: 347.87
Odchylenie standardowe liczby wystąpień jednego słowa: 230.82
```

Najpopularniejsze słowa



Długość słów

Średnia długość słowa: 7.35
Odchylenie standardowe: 2.46

Długość słów

- Zaciekawiły nas słówka bardzo krótkie i bardzo długie.

	word	nchars	occurences
1290	xi	2	1.0
1729	al	2	2.0
1959	iv	2	1.0
2500	ie	2	2.0
2578	ii	2	1.0
2967	st	2	2.0
3392	lo	2	2.0
3654	li	2	1.0
4204	nt	2	11.0
4356	ex	2	1.0
5185	ye	2	36.0
5348	ox	2	2.0
5982	om	2	2.0
5984	th	2	1.0
6018	em	2	1.0

	word	nchars	occurences
1	noseconsciousness	17	2.0
2	neitherpainfulnorpleasant	25	6.0
3	contradistinction	17	1.0
4	clingingsustenance	18	8.0
5	clingingaggregate	17	29.0
6	selfcomprehension	17	1.0
7	consciousnessconsciousness	26	1.0
8	neitherpleasurenorpain	22	2.0
9	bodyconsciousness	17	2.0
10	argumentativethought	20	1.0
11	stressfulsariputta	18	1.0
12	selfconsciousness	17	5.0
13	fabricationverbal	17	1.0
14	fabricationsfabrication	23	1.0
15	consciousnesshood	17	1.0
16	intellectconsciousness	22	2.0
17	neitherpleasantnorpainful	25	3.0
18	lamentationlamentation	22	1.0
19	soulconsciousness	17	1.0
20	tongueconsciousness	19	2.0
21	propertysariputta	17	1.0
22	clingingclingingsustenance	26	1.0
23	fabricationsmental	18	1.0

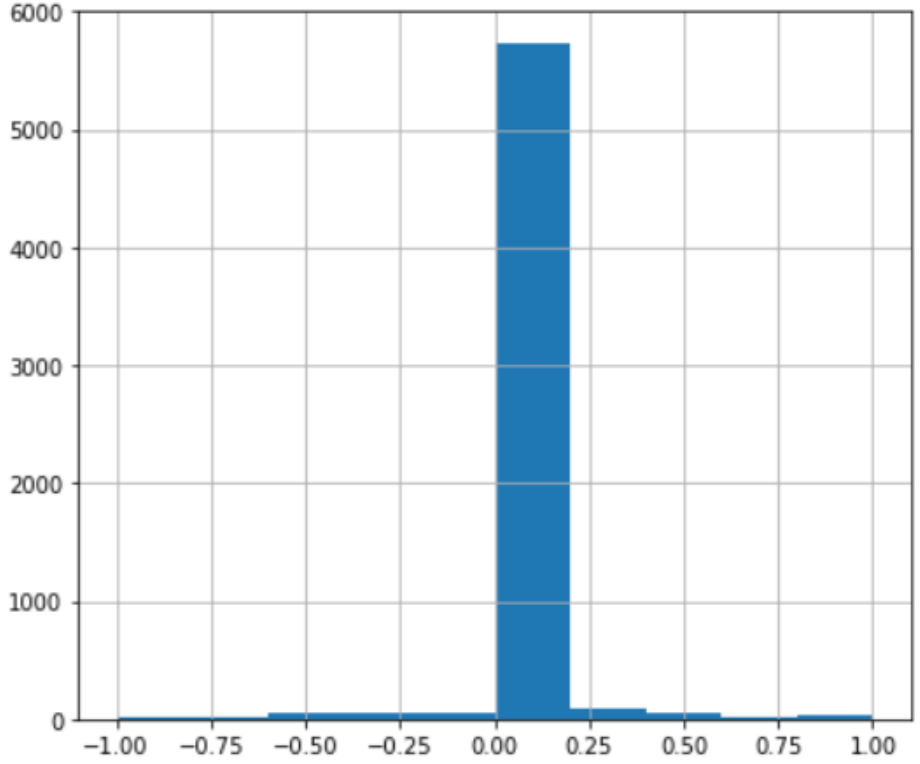
Długość słów

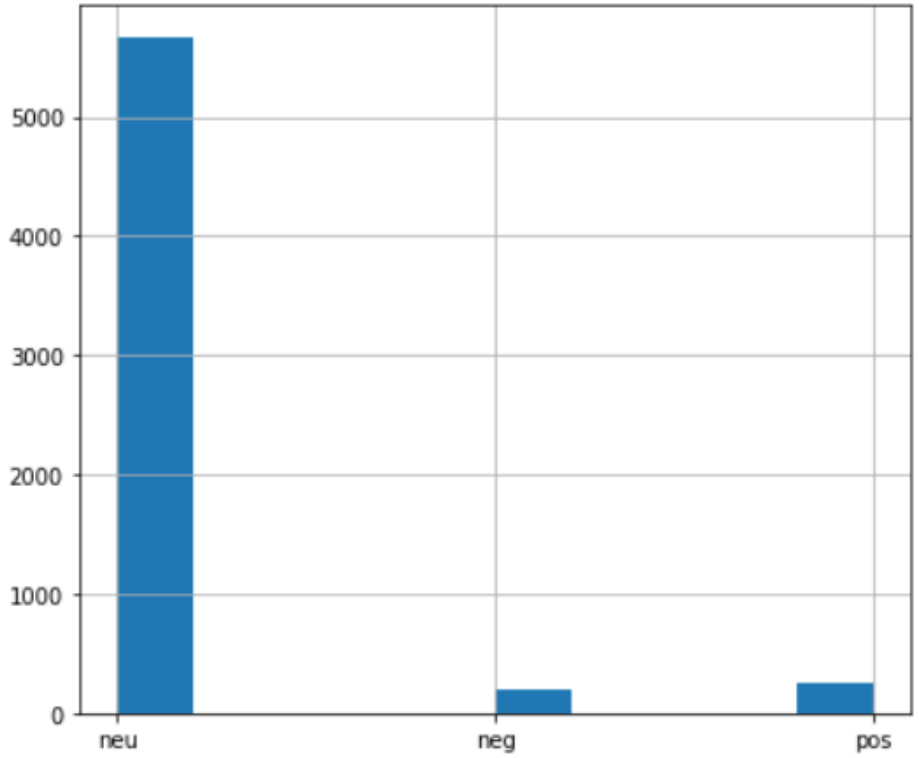
- Ponieważ ich liczba wystąpień jest bardzo mała (odchylenie standardowe równe 230), zdecydowaliśmy się te słowa również usunąć.

```
Ostateczny rozmiar ramki: (590, 6073).
```

- Ostatnim punktem naszego EDA było zbadanie nacechowania emocjonalnego naszego słownictwa. Wykorzystaliśmy do tego bibliotekę TextBlob.

Nacechowanie emocjonalne





- Na lewym wykresie, im wartość bliższa -1, tym słowo jest bardziej negatywne. Wartość 0 oznacza słowo neutralne. Wartości dodatnie - słowo pozytywne.
- Widzimy więc, że słowa w tekstach bibilijnych są głównie neutralne. Jest jednak minimalnie więcej pozytywnych niż negatywnych.