



Czerwiec 2023

NLP PROJECT

Projekt klasteryzacji

Magdalena Jeczeń
Marta Szuwarska

OPIS ZBIORU DANYCH

Zbiór danych **20 Newsgroups** to kolekcja około 20 000 dokumentów z grup dyskusyjnych, podzielonych (prawie) równomiernie na 20 różnych grup dyskusyjnych:

- sci.electronics
- talk.politics.mideast
- misc.forsale
- rec.sport.hockey
- comp.windows.x
- comp.sys.ibm.pc.hardware
- rec.sport.baseball
- comp.sys.mac.hardware
- sci.crypt
- talk.politics.misc
- comp.graphics
- comp.os.ms-windows.misc
- talk.politics.guns
- talk.religion.misc
- sci.space
- sci.med
- alt.atheism
- soc.religion.christian
- rec.motorcycles
- rec.autos

PREPROCESSING

Usunięcie nagłówków,

Sprawdzenie braków danych,

Czyszczenie danych (linki, adresy e-mail, znaki nowej linii itp.),

Rozpoznanie nazw własnych,

Zamiana dużych liter na małe,

Usunięcie tzw. stopwords

PREPROCESSING

Rozpoznanie części mowy,

Lematyzacja,

Grupowanie podobnych słów,

Sprawdzenie różnorodności leksykalnej, średniej długości słowa itp.,

Analiza sentymentalna,

Skalowanie i korelacja dla dodanych kolumn numerycznych.

OKREŚLENIE LICZBY KLASTRÓW



MODEL

Do klasteryzacji używaćmy metody KMeans.

Sprawdziłyśmy zależnie od liczby klastrów:

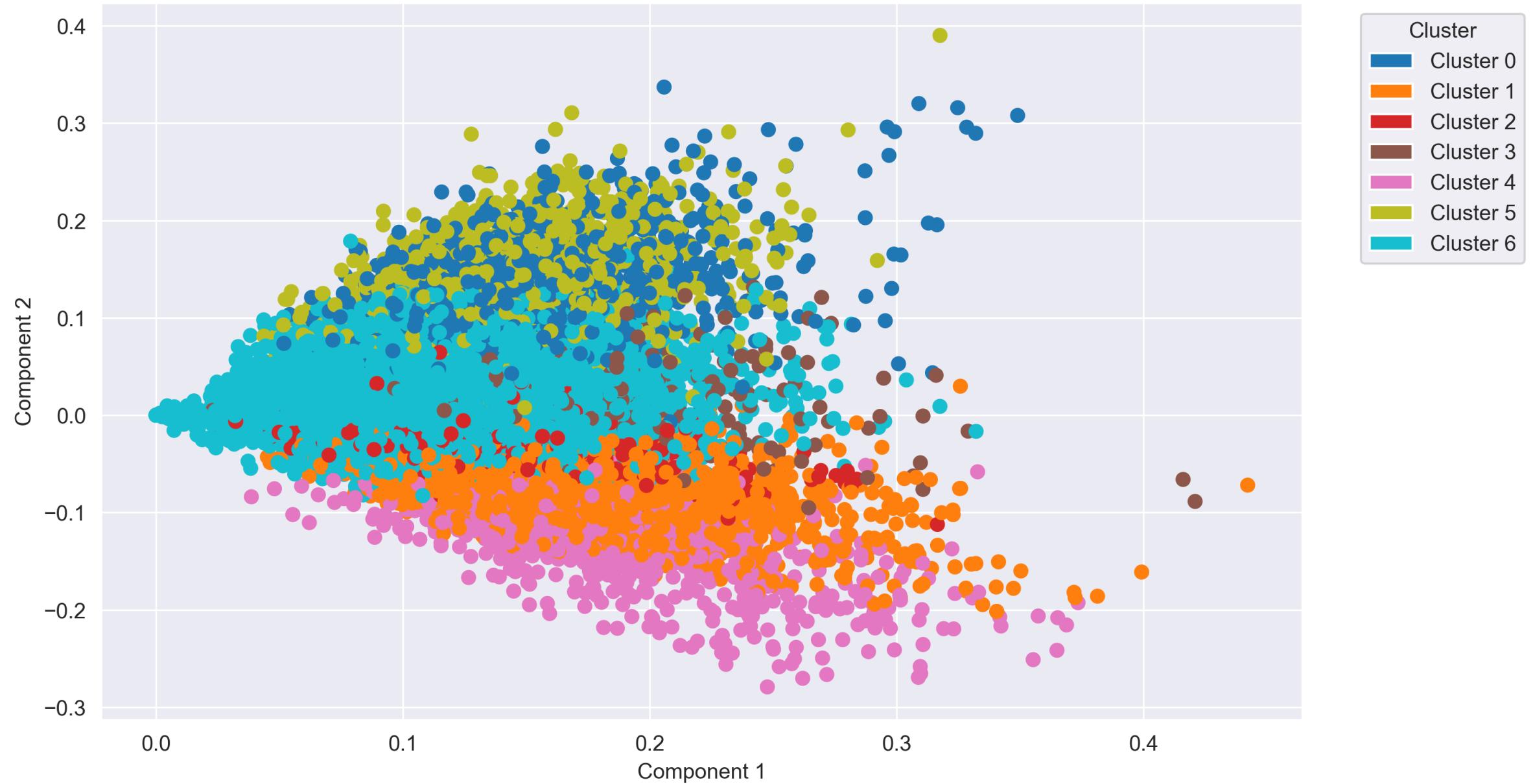
- Jak wygląda wykres klastrów po redukcji wymiarów,
- Jakie są najczęstsze słowa w danym klastrze,
- Jaki % rekordów z danej newsgroupy należy do danego klastra.

Na podstawie tych kryteriów wybrałyśmy model, który posiada 7 klastrów.



WYBRANY MODEL – 7 KLASTRÓW

Cluster Visualization



LICZBA OBSERWACJI
W KAŻDYM
Z KLASTRÓW

Cluster	Count
0	1246
1	2384
2	923
3	399
4	806
5	956
6	7252

NAJCZĘSTSZE SŁOWA W KLASTRZE

Cluster 0: window file program use run thanks image windows display application

Cluster 1: people say article right write government dont think gun make

Cluster 2: game team player play year hockey win fan season baseball

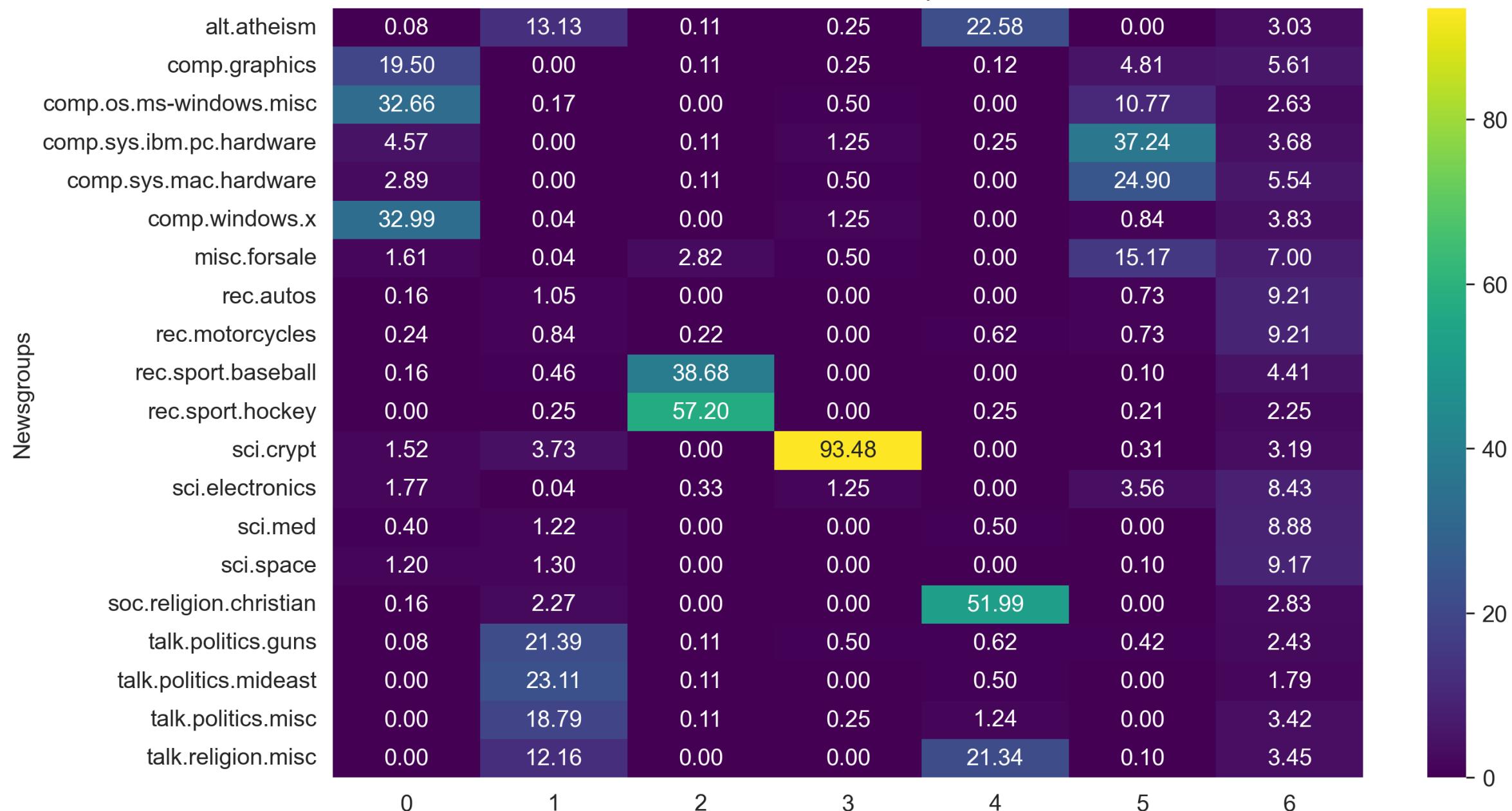
Cluster 3: key chip encryption clipper government use escrow phone algorithm nsa

Cluster 4: god jesus christian believe say bible christ people sin think

Cluster 5: drive card monitor scsi disk driver use problem controller video

Cluster 6: write article know use like thanks im car dont email

Ratio Heatmap



Cluster 0 – COMPUTER SOFTWARE

Cluster 1 - POLITICS

israel believe writes opinion country use batf
kill want fbi really gun fact
people good thing know moral
like claim start weapon
right article happen war
objective value mean koresh
government

Cluster 2 - HOCKEY AND BASEBALL

A word cloud visualization for Cluster 2, centered around Hockey and Baseball. The words are colored in various shades of green, blue, and purple, and are arranged in a roughly circular pattern. The most prominent words are 'game' (purple), 'player' (green), 'baseball' (green), 'team' (yellow-green), 'hockey' (yellow-green), and 'article' (blue). Other visible words include 'win', 'ranger', 'good', 'im.', 'coach', 'division', 'cup', 'guy', 'espn', 'best', 'boston', 'pitch', 'fan', 'series', 'nhl', 'hit', 'blue', 'league', 'think', 'year', 'really', 'dont', 'season', 'play', 'know', 'lose', 'say', 'score', 'wing', 'come', 'goal', 'night', 'detroit', 'make', 'pitcher', 'run', 'tradetime', 'brave', 'playoff', 'great', and 'article'.

Cluster 3 - CRYPTOGRAPHY

enforcement court scheme crypto company
david phone clipper use law
communication escrow nsa write
data algorithm tap security need
sternlight encryption number
private message public information agency people
secure encrypt say
key article wiretap pgp chip
secret proposal serial dont make trust

Cluster 4 - RELIGION

people think make book bible believe atheist article sin jesus hell religion

christianity

truth know say tell life way

point word question

mean come

like accept

time claim

church

scripture

man

die mary evidence

love

Cluster 5 – COMPUTER HARDWARE

Cluster 6 – BASIC DISCUSSION

problem sell people
like make new
need question buy
want phone university
work think year time
use good write
good day book tell
space information send
list help
send bike
say articles
read really come
impost
give things
ask look price mail
try dont car
writes

UWAGI WALIDACJI

- Dodanie interpretacji klastrów - uwzględnione,
- Sprawdzenie innych modeli - pominięte ze względu na wystarczająco dobre wyniki z użyciem KMeans,
- Podział na zbiór treningowy i testowy - uwzględnione (dodana cross-walidacja).

PODSUMOWANIE



Nasz model podzielił artykuły z 20 Newsgroups na 7 klastrów.



Połączył zestawy z podobnych newsgroups.

**DZIĘKUJEMY
ZA UWAGĘ**

