The 20 Newsgroups data set clustering

Michał Iwicki
Mateusz Nizwantowski

Zbiór danych

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

Przykładowy mail

Xref: cantaloupe.srv.cs.cmu.edu talk.abortion:121624 alt.atheism:54236 talk.religion.misc:84407
Newsgroups: talk.abortion,alt.atheism,talk.religion.misc
Path:
cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!fs7.ece.cmu.edu!europa.eng.gtefsd.com!howland.reston.ans.net!zaphod.mps.ohio-state.edu!cs.utexas.edu!asuvax!ncar!noao!forgach
From: forgach@noao.edu (Suzanne Forgach)
Subject: Re: After 2000 years, can we say that Christian Morality is
Message-ID: <1993Apr23.005217.11121@noao.edu>
Originator: forgach@gemini.tuc.noao.edu
Sender: news@noao.edu
Nntp-Posting-Host: gemini.tuc.noao.edu
Organization: National Optical Astronomy Observatories, Tucson, AZ, USA
References: <1993Apr16.041641.22140@leland.Stanford.EDU>
Date: Fri, 23 Apr 1993 00:52:17 GMT
Lines: 5

> In article <kmr4.1587.734911207@po.CWRU.edu> kmr4@po.CWRU.edu (Keith M. Ryan) writes:
>
> Only when the Sun starts to orbit the Earth will I accept the Bible.

Did you forget that two spinning skaters are in orbit around each other?

Business case

- Odkryć tematy rozmów, aby przyporządkować istniejących użytkowników do nowych grup wyznaczonych względem tematów rozmów
- Pomoc w odnajdywaniu rozmówców o podobnych zainteresowaniach
- Personalizacja treści

Preprocessing

- Ograniczenie do samej treści maila
- Pozostawienie samych liter
- Tokenizacja
- Usunięcie stop wordsów
- Lematyzacja plus stemming

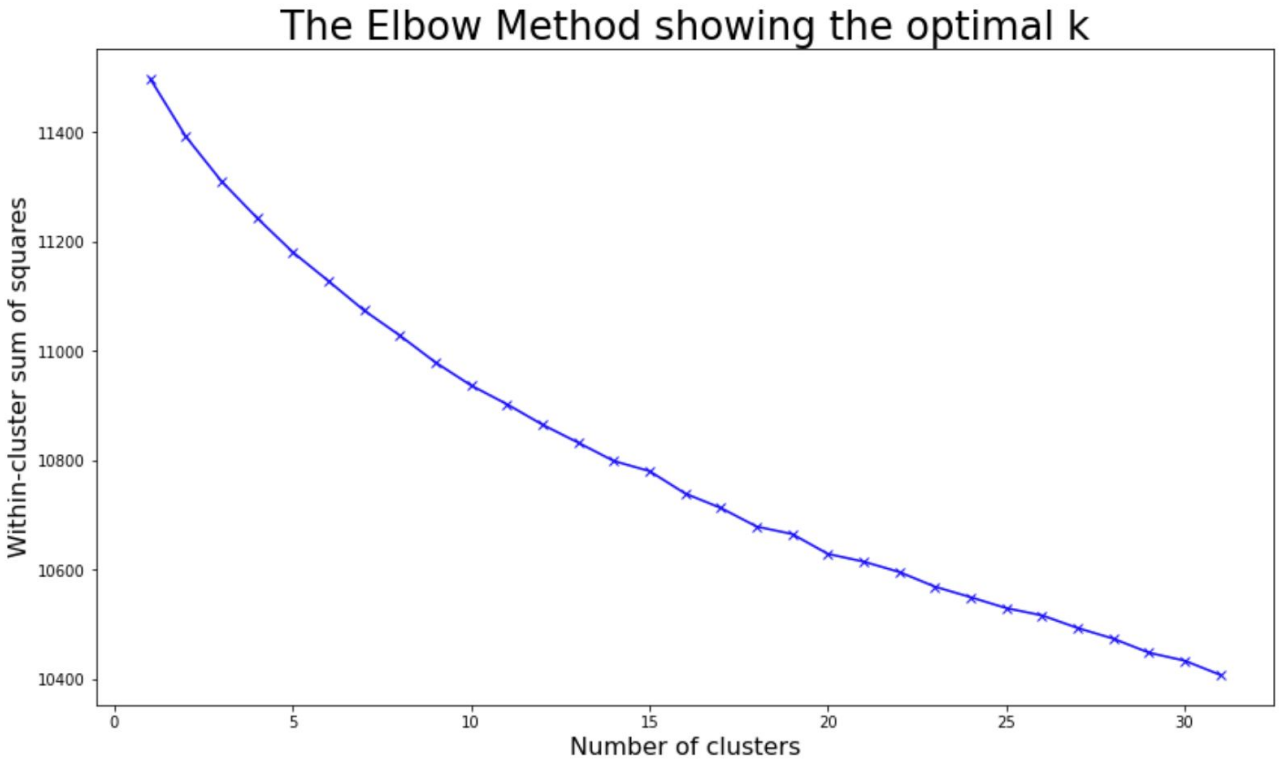
Wektoryzacja z Tfidf

	abil	abl	absolut	accept	access	accord	account	act	action	activ	...	world	wors	worth	written	wrong	wrote	ye	year	york
0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.190352	0.0
1	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.232458	0.000000	0.0
2	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
3	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
4	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11993	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.193387	0.000000	0.0
11994	0.0	0.0	0.0	0.105819	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.134014	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
11995	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
11996	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
11997	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.062128	0.0

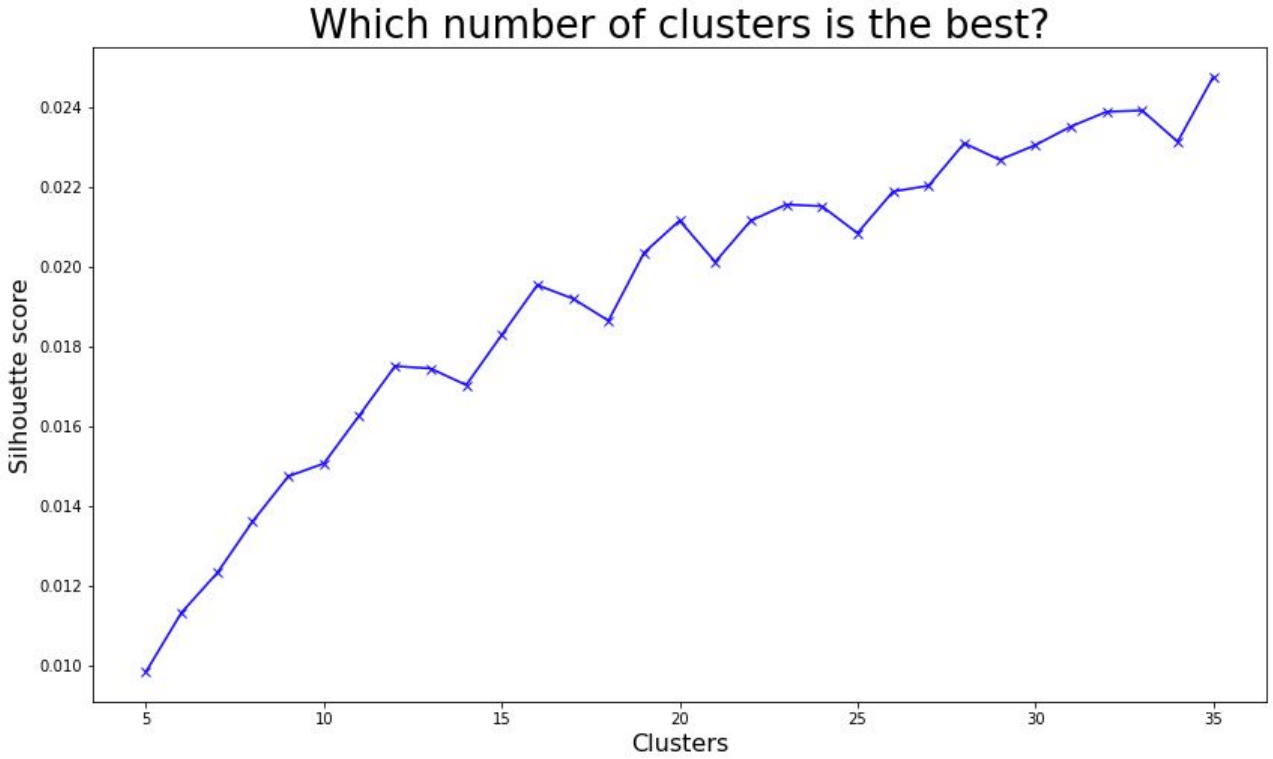
11998 rows × 797 columns

Usunęliśmy słowa występujące częściej niż w 20% tekstów i rzadziej niż w 2%

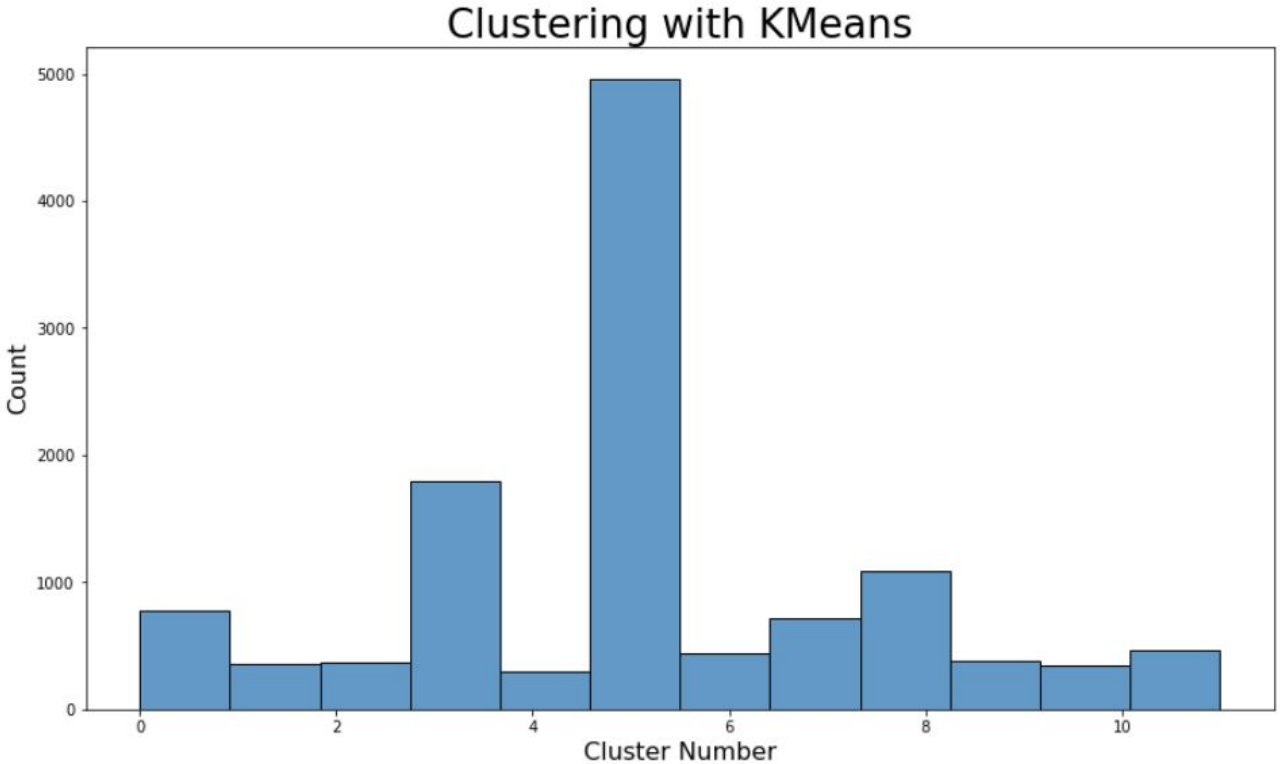
Określenie ile potrzeba klastrów do Kmeans



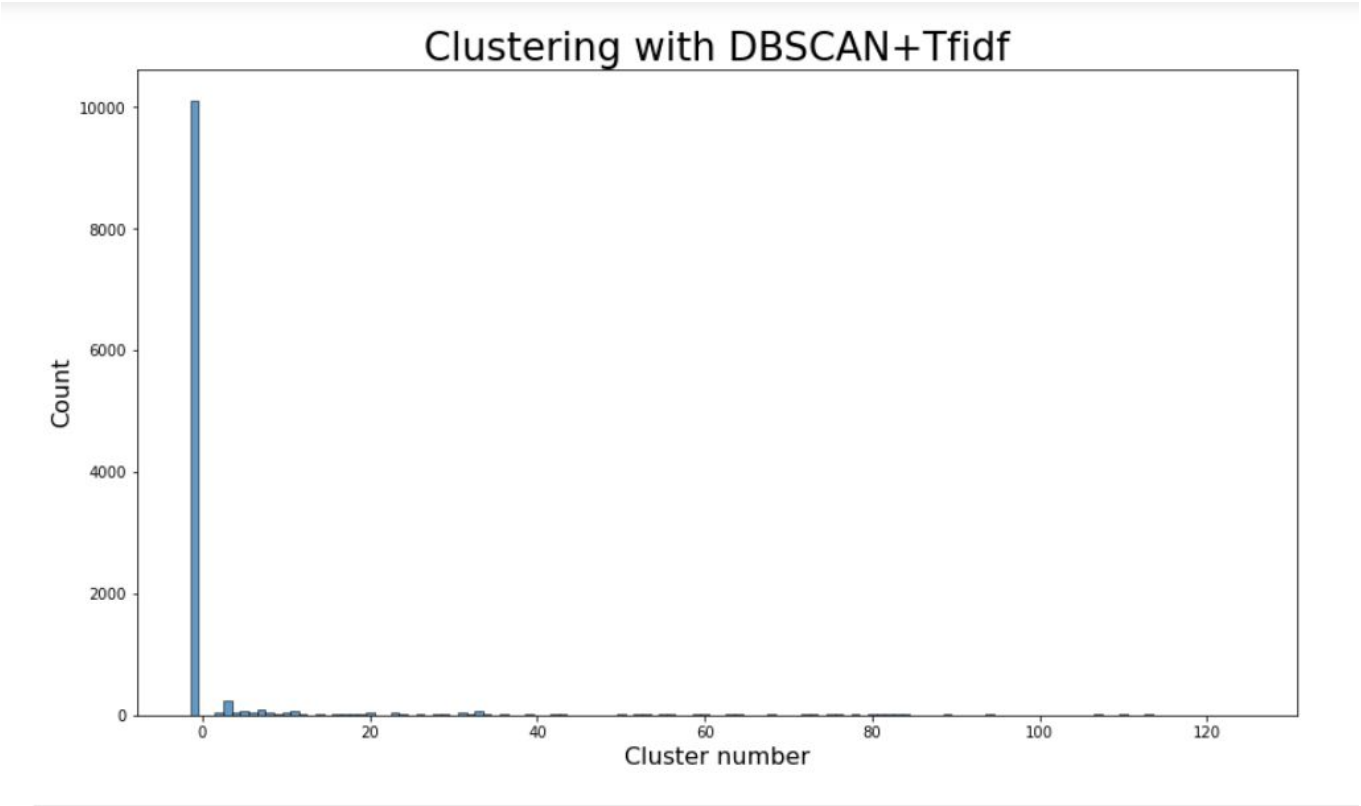
Silhouette score dla KMeans



Rozkład KMeans



Rozkład DBSCAN

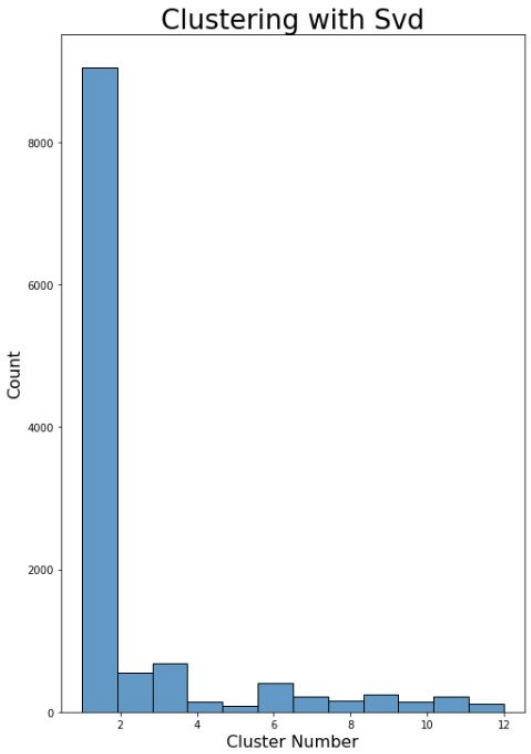


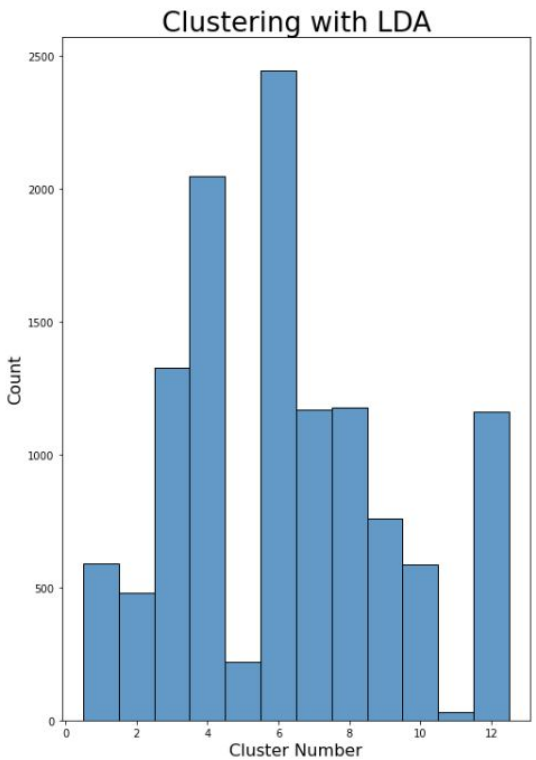
Zmiana podejścia

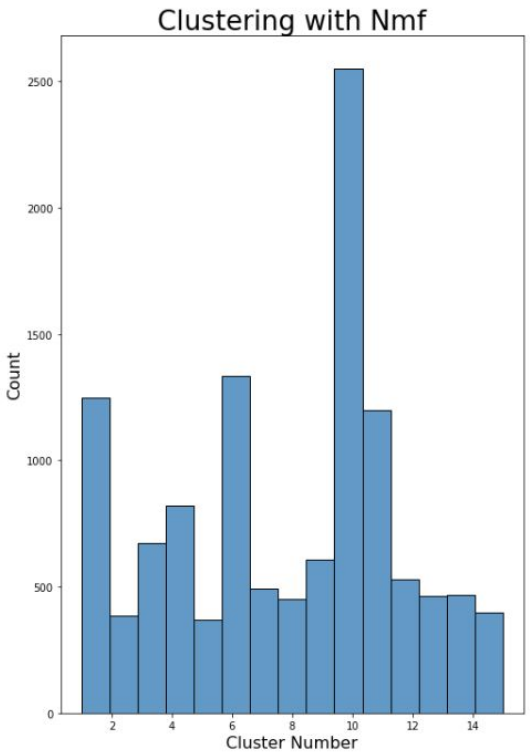
Klasyczne metody klasteringu widzą dane jako duży szum. Zatem zmieniamy podejście i się skupimy na redukcji wymiarów i klastrowaniu wokół nowych kolumn.

Porównamy metody rozkładu SVD, LDA i NMF, żeby znaleźć ten który najlepiej radzi sobie z tym problemem.

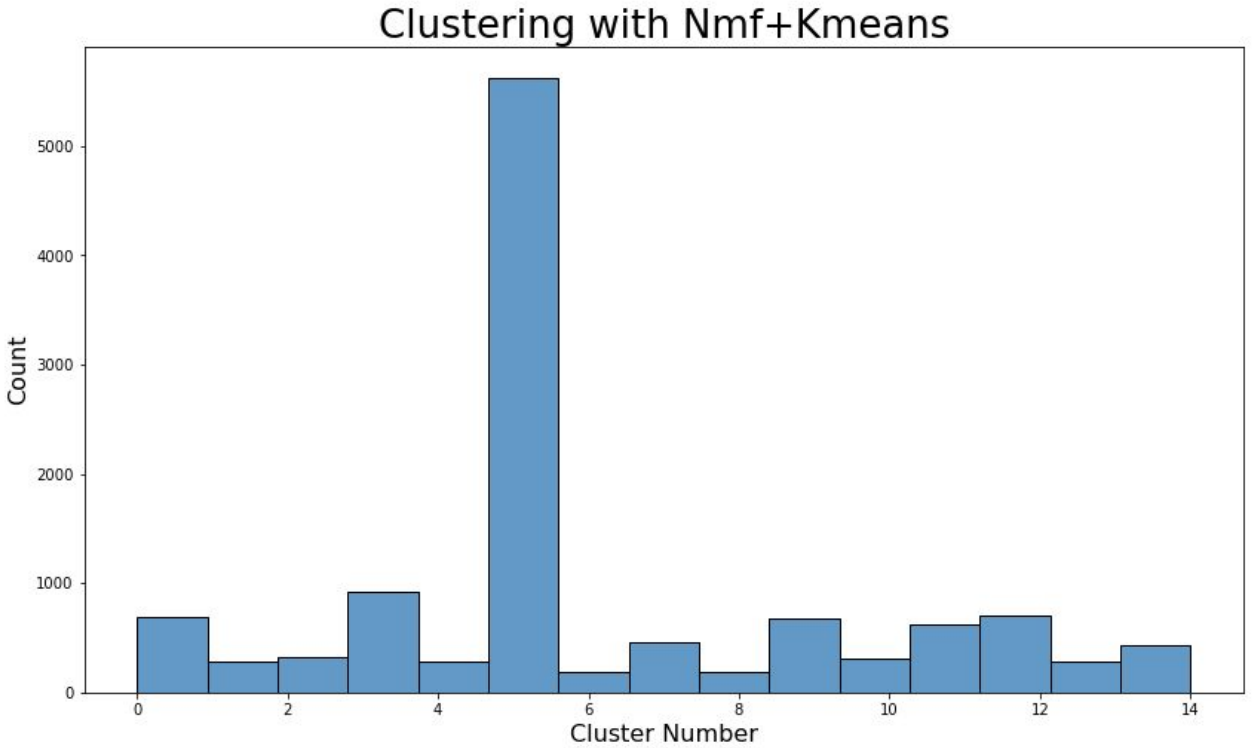
SVD vs LDA vs NMF



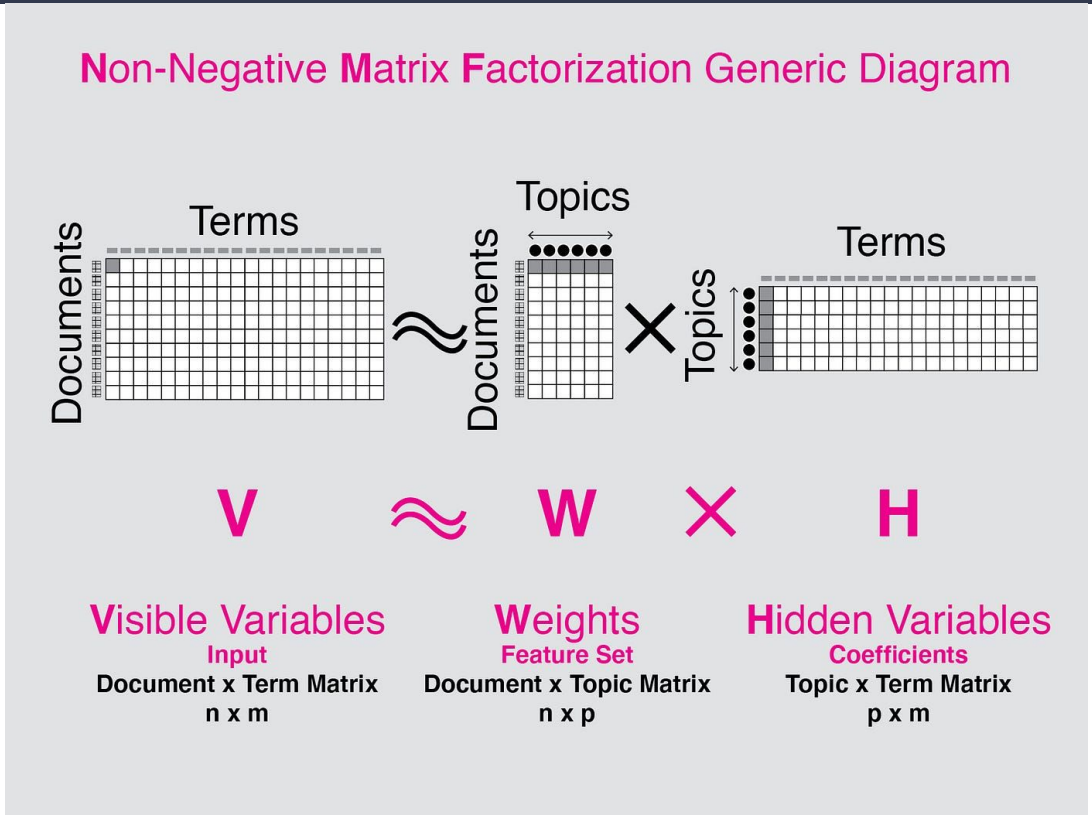




Czy warto zrobić clustering po NMF?



Czym jest NMF i jak nim klastrowaliśmy?

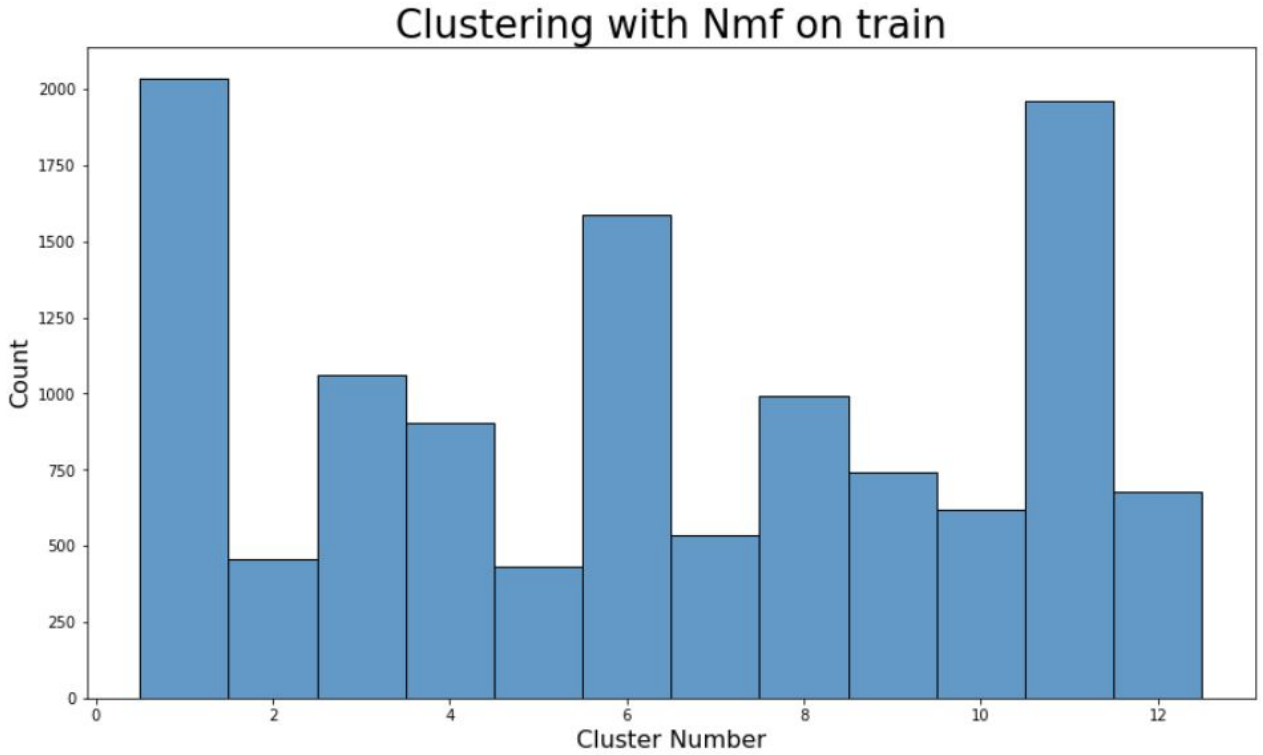


Określenie ile potrzeba klastrów

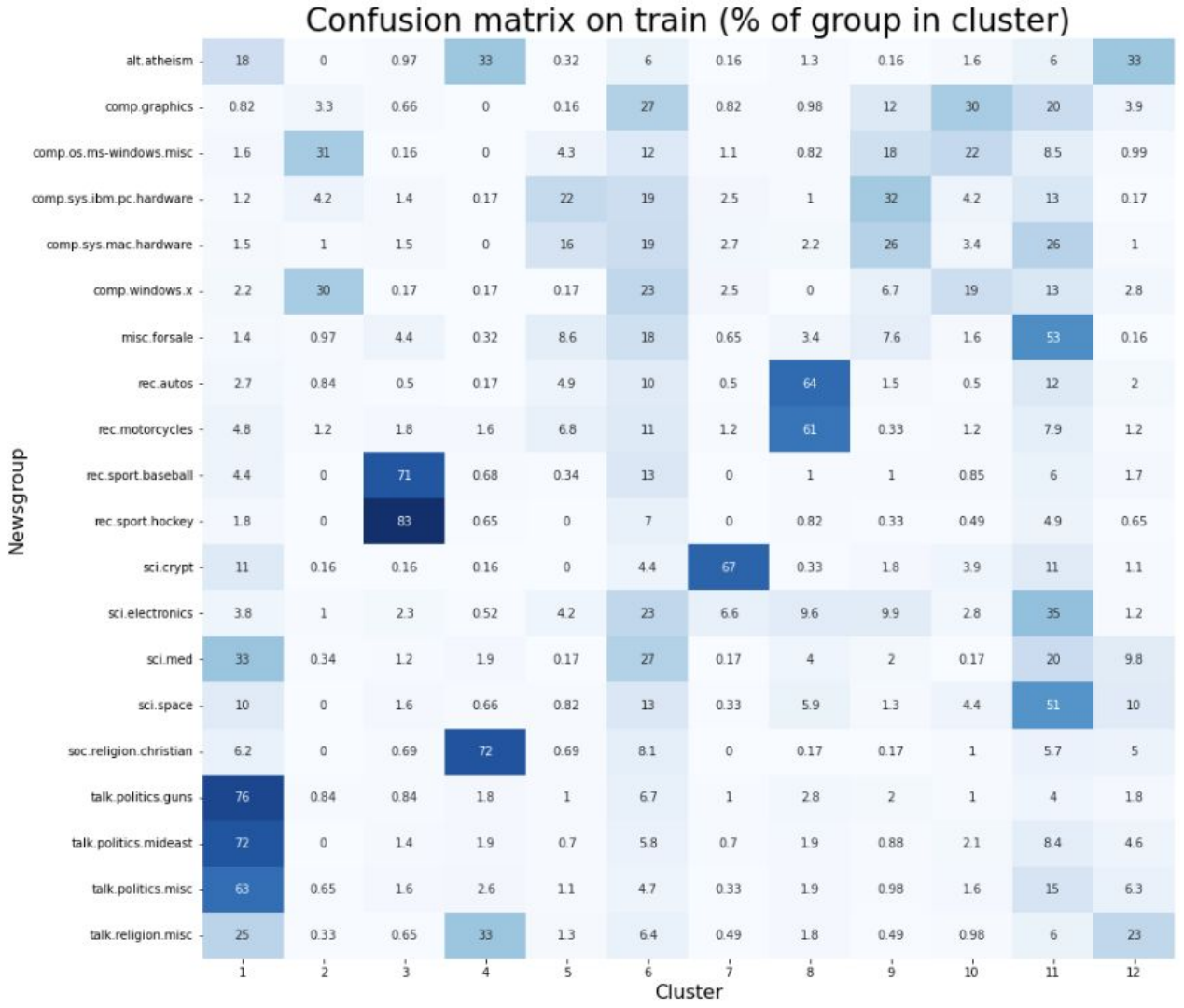
Jakie tematy zostały wygenerowane?

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Topic # 01	right	govern	gun	state	law	kill	said	child	fbi	way	israel	want	thing	year	day
Topic # 02	window	run	program	problem	applic	manag	version	display	driver	screen	user	instal	set	machin	work
Topic # 03	game	team	play	player	year	win	fan	season	basebal	leagu	hit	goal	good	run	watch
Topic # 04	god	christian	jesu	believ	bibl	christ	faith	exist	church	love	belief	life	religion	word	man
Topic # 05	drive	disk	hard	mb	control	problem	work	mac	instal	format	set	speed	power	switch	need
Topic # 06	thank	pleas	mail	anyon	advanc	post	email	appreci	hi	repli	help	send	address	list	look
Topic # 07	key	chip	encrypt	clipper	secur	bit	phone	public	govern	number	messag	data	devic	law	privat
Topic # 08	car	bike	engin	ride	dod	speed	look	good	buy	mile	light	road	realli	thing	got
Topic # 09	card	driver	video	monitor	bit	color	mb	problem	board	work	graphic	mode	control	ram	pc
Topic # 10	file	imag	format	program	convert	ftp	display	graphic	disk	color	tri	avail	site	read	creat
Topic # 11	new	price	comput	univers	sale	space	offer	includ	sell	softwar	book	phone	work	ship	box
Topic # 12	object	moral	valu	frank	scienc	mean	uucp	theori	observ	better	good	differ	exist	subject	truth

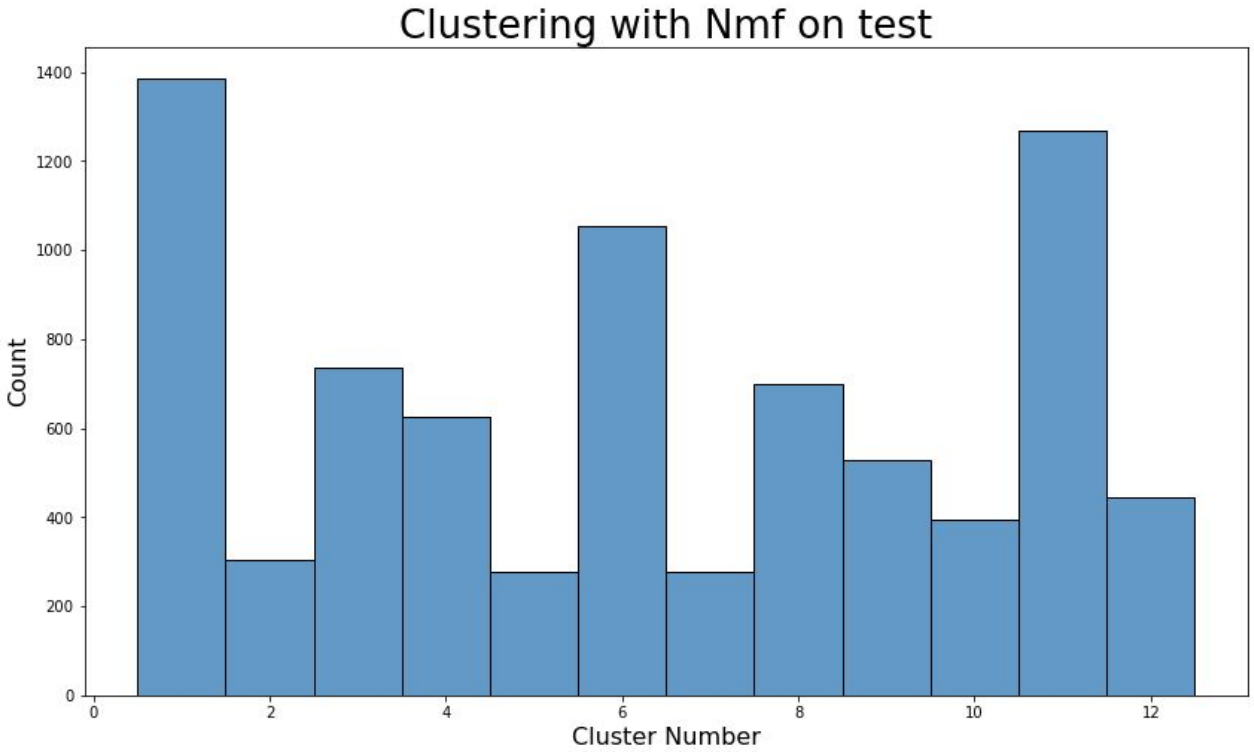
Rozkład dla najlepszej liczby klastrów



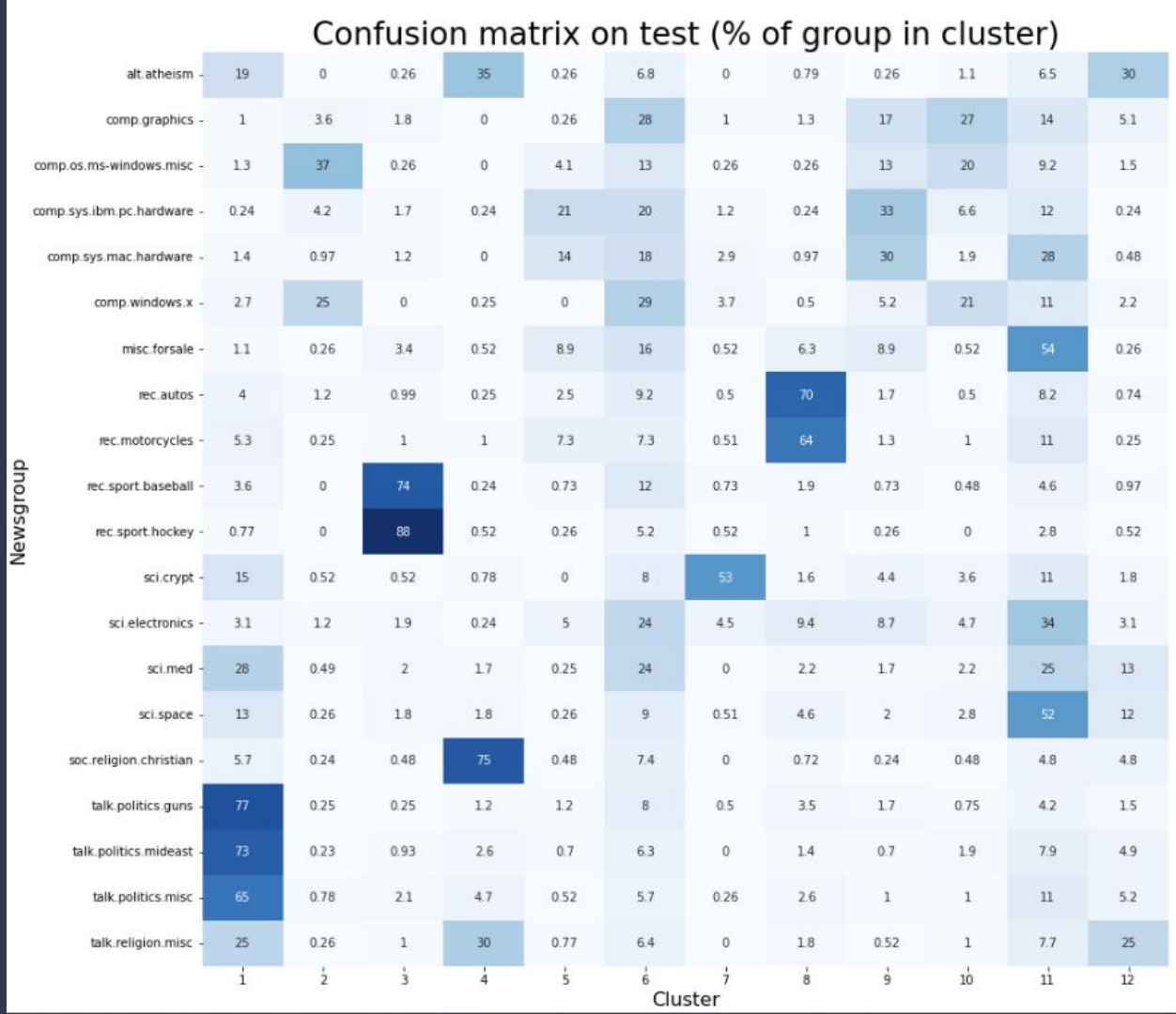
Jak one się pokrywają z oryginalnymi grupami?



Rozkład na zbiorze testowym

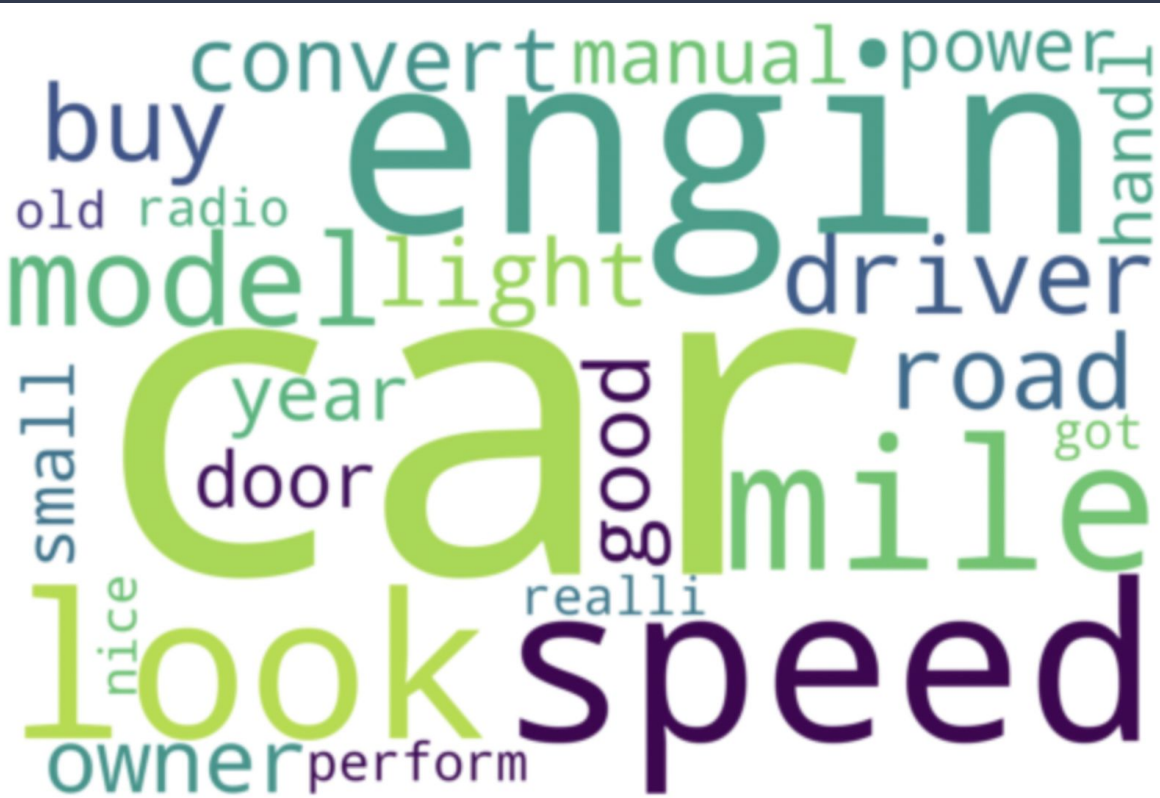


Zgodność na zbiorze testowym z oryginalnym podziałem



Israeli-Palestinian Conflict and Politics

Automotive and Driving Experience

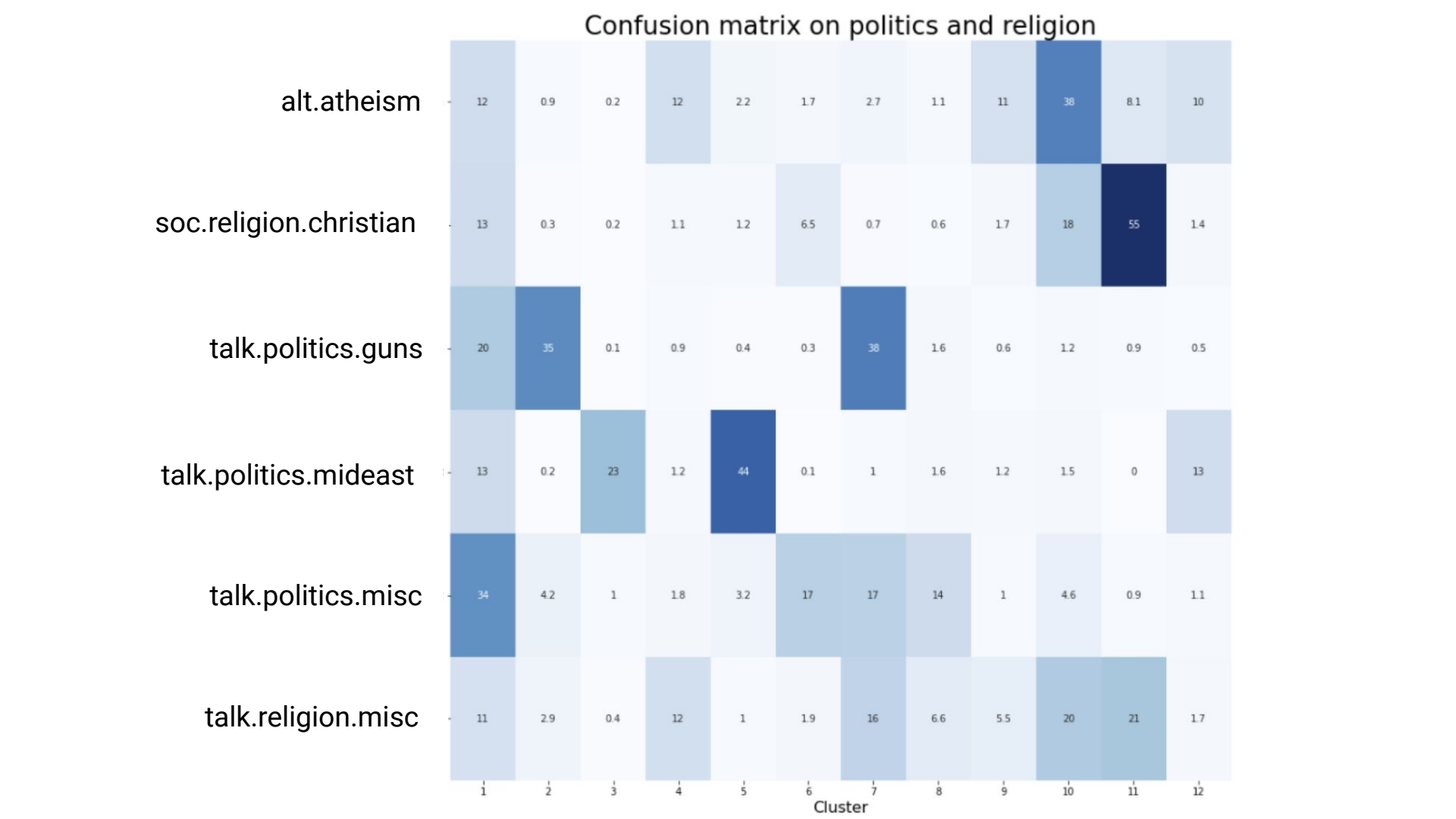


Digital Security and Encryption

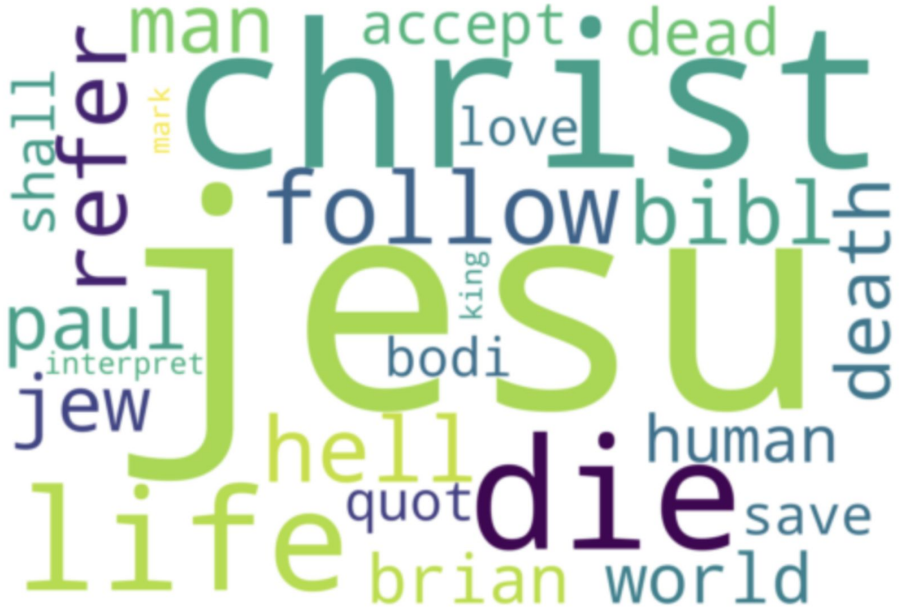
Ciekawe wyniki na ograniczonym zbiorze danych

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Topic # 01	post	govern	year	presid	new	american	book	read	univers	clinton	pleas	mail	group	nation	com
Topic # 02	gun	law	weapon	firearm	crimin	crime	control	polic	arm	kill	shoot	carri	safeti	car	death
Topic # 03	armenian	turkish	armenia	turk	soviet	genocid	turkey	extermin	mountain	villag	serdar	argic	road	escap	massacr
Topic # 04	valu	object	frank	dwyer	scienc	uucp	observ	better	realiti	truth	subject	eric	predict	deal	judgement
Topic # 05	israel	isra	arab	jew	palestinian	jewish	kill	peac	attack	occupi	territori	civilian	land	soldier	polici
Topic # 06	homosexu	cramer	gay	clayton	men	sexual	com	sex	male	number	uunet	studi	relat	straight	child
Topic # 07	fbi	koresh	batf	bd	compound	start	ga	child	atf	davidian	cult	tank	burn	agent	claim
Topic # 08	abort	insur	pay	health	tax	privat	care	fund	money	reduc	servic	cost	program	spend	larri
Topic # 09	moral	object	absolut	jon	action	forc	keith	wrong	human	societi	natur	law	murder	anim	behavior
Topic # 10	exist	atheist	theori	religion	belief	atheism	scienc	evid	faith	argument	reason	creation	claim	truth	true
Topic # 11	christian	jesu	christ	sin	church	bibl	love	faith	heaven	scriptur	life	lord	word	book	die
Topic # 12	muslim	islam	jew	religion	war	religi	ethnic	nazi	world	jewish	edu	law	genocid	christian	countri

Ograniczyliśmy się do grup związanych z polityką i religią


Confusion matrix on politics and religion
alt.atheism
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc
12 0.9 0.2 12 2.2 1.7 2.7 1.1 11 38 8.1 10
13 0.3 0.2 1.1 1.2 6.5 0.7 0.6 1.7 18 55 1.4
20 35 0.1 0.9 0.4 0.3 38 1.6 0.6 1.2 0.9 0.5
13 0.2 23 1.2 44 0.1 1 1.6 1.2 1.5 0 13
34 4.2 1 1.8 3.2 17 17 14 1 4.6 0.9 1.1
11 2.9 0.4 12 1 1.9 16 6.6 5.5 20 21 1.7
1 2 3 4 5 6 7 8 9 10 11 12
Cluster

A co się stanie gdy drastycznie zwiększymy liczbę klastrów?



Współpraca z zespołem walidacyjnym

Była bardzo owocna i zespół walidacyjny nam bardzo dużo pomógł. Odwołaliśmy się do ich uwag i je poprawiliśmy między innymi:

- Jaśniejsze i więcej znaczące nazwy funkcji
- Językowa spójność komentarzy (ujednolicenie języka)
- Utwierdzili nas w przekonaniu że nasza metodyka jest skuteczna i poprawna
- Dzięki nim zdaliśmy sobie sprawę, że to nie błąd jeżeli jeden klaster jest większy niż pozostałe

Ich doświadczenie z NLP nabyte podczas pierwszego projektu było nieocenione i pomogło nam znacząco.