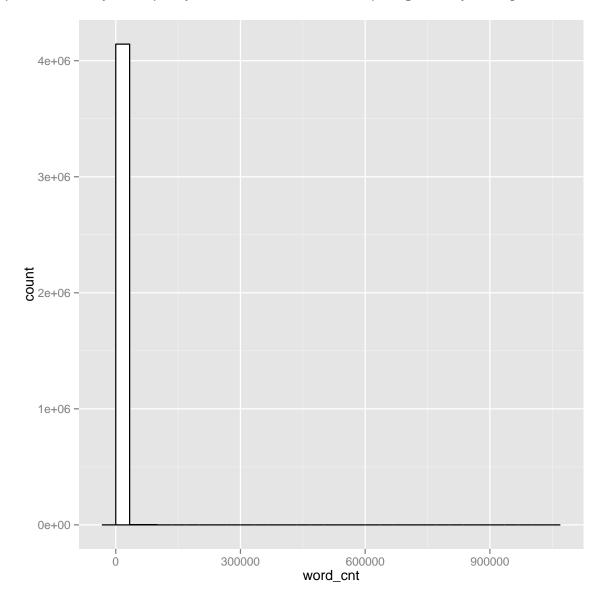
Słowa - statystyki

Natalia Potocka

29 grudnia 2014

W polskiej Wikipedii znaleziono 4143353 unikalnych słów. 55.05% z nich występuje dokładnie tylko w jednym tekście. Dobrze obrazuje to poniższy histogram.



Słowa występujące tylko w jednym tekście to zazwyczaj liczby oraz słowa w obcych językach, przykładowo:

[1]	"interaktiver"	"gloveria"	"d'axa"
[4]	"pl516"	"jahrhundertbrücke"	"111860"
[7]	"kundelinga"	"bajaprog"	"konkretniejsza"
[10]	"julgneville"		

TT 1 1 1 T 1 (1	. ,		1 1/
Tabela 1: Lista	_dziesieciii	naiczesciei	l wysteniiia <i>c</i>	vch słow
Tubcia I. Libia	azicoiçcia	TIMICECUCICI	vv y btcp alac	y CIL DIO VV.

Słowo	Liczba artykułów		
W	1002731		
i	728424		
na	690433		
Z	652071		
do	561449		
się	536522		
roku	419538		
a	382204		
od	378804		
jest	342802		

Po usunięciu słów występujących w jednym, dwóch lub trzech tekstach otrzymujemy następujące statystyki:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4	[10	145	30	1000000

Słowa, które występują przynajmniej w czterech stanowią zaledwie 25.06% wszystkich słów.

Słowa występujące w największej liczbie tekstów to:

Są to słowa nie istotne w kontekście analizy tematycznej tekstu. Podobnych słów jest więcej. Ich pełna lista zostaje przedstawiona poniżej:

[1]	"ach"	" aj "	"albo"	"bardzo"	"bez"
[6]	"od"	"być"	"ci"	"cię"	"ciebie"
[11]	"co"	"CZY"	"daleko"	"dla"	"dlaczego'
[16]	"dlatego"	"do"	"dobrze"	"dokąd"	"dość"
[21]	"dużo"	"dwa"	"dwaj"	"dwie"	"dwoje"
[26]	"dziś"	"dzisiaj"	"gdyby"	"gdzie"	"go"

[31]	"ich"	"ile"	"im"	"inny"	"ja"
[36]	"ją"	"jak"	"jakby"	"jaki"	"je"
[41]	"jeden"	"jedna"	"jedno"	"jego"	"jej"
[46]	"jemu"	"jeśli"	"jest"	"jestem"	"jeżeli"
[51]	"już"	"każdy"	"kiedy"	"kierunku"	"kto"
[56]	"ku"	"lub"	"ma"	"mają"	"mam"
[61]	"mi"	"mną"	"mnie"	"moi"	"mój"
[66]	"moja"	"moje"	"może"	"mu"	"my"
[71]	"na"	"nam"	"nami"	"nas"	"nasi"
[76]	"nasz"	"nasza"	"nasze"	"natychmiast"	"nią"
[81]	"nic"	"nich"	"nie"	"niego"	"niej"
[86]	"niemu"	"nigdy"	"nim"	"nimi"	"niż"
[91]	"obok"	"od"	"około"	"on"	"ona"
[96]	"one"	"oni"	"ono"	"owszem"	"po"
[101]	"pod"	"ponieważ"	"przed"	"przedtem"	"są"
[106]	"sam"	"sama"	"się"	"skąd"	"tak"
[111]	"taki"	"tam"	"ten"	"to"	"toba"
[116]	"tobie"	"tu"	"tutaj"	"twoi"	"twój"
[121]	"twoja"	"twoje"	"ty"	"wam"	"wami"
[126]	"was"	"wasi"	"wasz"	"wasza"	"wasze"
[131]	"we"	"więc"	"wszystko"	"wtedy"	"wy"
[136]	"żaden"	"zawsze"	"że"		
	[36] [41] [46] [51] [56] [61] [66] [71] [76] [81] [86] [91] [96] [101] [106] [111] [116] [121] [126] [131]	[36] "ja" [41] "jeden" [46] "jemu" [51] "już" [56] "ku" [61] "mi" [66] "moja" [71] "na" [76] "nasz" [81] "nic" [86] "niemu" [91] "obok" [96] "one" [101] "pod" [106] "sam" [111] "taki" [116] "tobie" [121] "twoja" [126] "was" [131] "we"	[36] "ja" "jak" [41] "jeden" "jedna" [46] "jemu" "jeśli" [51] "już" "każdy" [56] "ku" "lub" [61] "mi" "mna" [66] "moja" "moje" [71] "na" "nam" [76] "nasz" "nasza" [81] "nic" "nich" [86] "niemu" "nigdy" [91] "obok" "od" [96] "one" "oni" [101] "pod" "ponieważ" [106] "sam" "sama" [111] "taki" "tam" [116] "tobie" "tu" [121] "twoja" "twoje" [126] "was" "wasi" [131] "we" "więc"	[36] "ja" "jak" "jakby" [41] "jeden" "jedna" "jedno" [46] "jemu" "jeśli" "jest" [51] "już" "każdy" "kiedy" [56] "ku" "lub" "ma" [61] "mi" "mna" "mnie" [66] "moja" "moje" "może" [71] "na" "nam" "nami" [76] "nasz" "nasza" "nasze" [81] "nic" "nich" "nie" [86] "niemu" "nigdy" "nim" [91] "obok" "od" "około" [96] "one" "oni" "ono" [101] "pod" "ponieważ" "przed" [106] "sam" "sama" "się" [111] "taki" "tam" "ten" [112] "twoja" "twoje" "ty" [126] "was" "wasz" "wasz" [131] "we" "więc" "wszystko"	[36] "ja" "jak" "jakby" "jaki" [41] "jeden" "jedna" "jedno" "jego" [46] "jemu" "jeśli" "jest" "jestem" [51] "już" "każdy" "kiedy" "kierunku" [56] "ku" "lub" "ma" "maja" [61] "mi" "mna" "może" "mu" [66] "moja" "moje" "może" "mu" [71] "na" "nam" "nasze" "natychmiast" [81] "nic" "nich" "nie" "nie" "niego" [88] "niemu" "nigdy" "nim" "nimi" [91] "obok" "od" "około" "on" [96] "one" "oni" "ono" "owszem" [101] "pod" "ponieważ" "przed" "przedtem" [106] "sam" "sama" "się" "skad" [111] "taki" "tam" "ten" "to" [116] "tobie" "tu" "tutaj" "twoi" [121] "twoja" "twoje" "ty" "wasza" [131] "we" "więc" "wszystko" "wtedy"