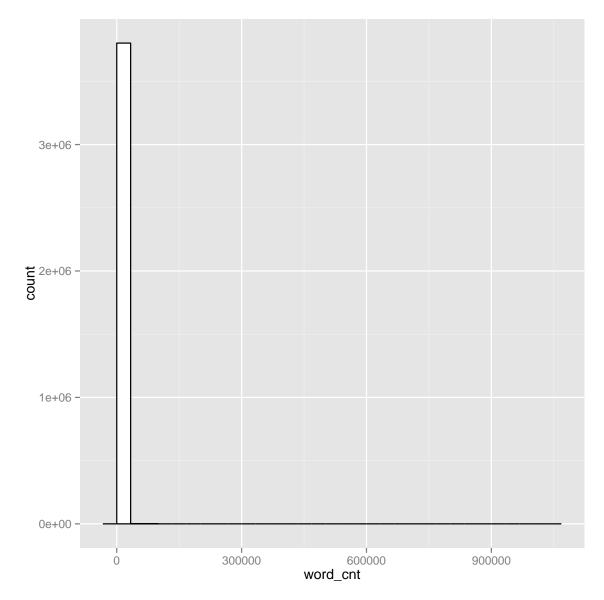
## Słowa - statystyki

Natalia Potocka

30 grudnia 2014

W polskiej Wikipedii znaleziono 3802968 unikalnych słów. 56.74% z nich występuje dokładnie w jednym tekście, natomiast 51.7% występuje tylko raz. Dobrze obrazuje to poniższy histogram.



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

[1] "300346"	"bua2"	"przechowujesz"	"168524"
[5] "bedoin"	"78341"	"2020551446"	"niewspomniani"
[9] "terdsakiem"	"anamorf"		

choć czasem są to słowa będące odmianą słów bardziej częściej występujących, np. słowo *przechowujesz* jest odmianą czasownika *przechowywać*. Należy wziąć pod uwagę takie słowa, gdyż mogą one polepszyć jakość dopasowania tekstów pod względem tematycznym.

Słowa występujące w największej liczbie tekstów to:

	Słowo	Liczba artykułów	Liczba wystąpień słowa
1	W	1001499	10246420
2	i	723899	4240658
3	na	687441	3183731
4	Z	647318	3218341
5	do	557602	2273872
6	się	535273	2066420
7	roku	418980	1284882
8	a	377247	908846
9	od	376645	926187
10	jest	342055	977374

W większości przypadków 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]	"bo"	"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"

"ku"	"lub"	"ma"	"mają"	"mam"
"mi"	"mną"	"mnie"	"moi"	"mój"
"moja"	"moje"	"może"	"mu"	"my"
"na"	"nam"	"nami"	"nas"	"nasi"
"nasz"	"nasza"	"nasze"	"natychmiast"	"nią"
"nic"	"nich"	"nie"	"niego"	"niej"
"niemu"	"nigdy"	"nim"	"nimi"	"niż"
"obok"	"od"	"około"	"on"	"ona"
"one"	"oni"	"ono"	"owszem"	"po"
"pod"	"ponieważ"	"przed"	"przedtem"	"są"
"sam"	"sama"	"się"	"skąd"	"tak"
"taki"	"tam"	"ten"	"to"	"toba"
"tobie"	"tu"	"tutaj"	"twoi"	"twój"
"twoja"	"twoje"	"ty"	"wam"	"wami"
"was"	"wasi"	"wasz"	"wasza"	"wasze"
"We"	"więc"	"wszystko"	"wtedy"	" <sub>WY</sub> "
"żaden"	"zawsze"	"że"		
	"mi" "moja" "nasz" "nic" "niemu" "obok" "one" "pod" "sam" "taki" "tobie" "twoja" "was" "we"	<pre>"mi" "mna" "moje" "na" "nam" "nasz" "nasza" "nic" "nich" "niemu" "nigdy" "obok" "od" "one" "oni" "pod" "ponieważ" "sam" "sama" "taki" "tam" "tobie" "tu" "twoja" "twoje" "was" "wasi" "we" "więc"</pre>	<pre>"mi" "mna" "mnie" "moja" "moje" "może" "na" "nam" "nami" "nasz" "nasza" "nasze" "nic" "nich" "nie" "niemu" "nigdy" "nim" "obok" "od" "około" "one" "oni" "ono" "pod" "ponieważ" "przed" "sam" "sama" "się" "taki" "tam" "ten" "tobie" "tu" "tutaj" "twoja" "twoje" "ty" "was" "wasi" "wasz" "we" "więc" "wszystko"</pre>	"mi"       "mna"       "moie"       "moi"         "moja"       "może"       "mu"         "na"       "nam"       "nami"       "nas"         "nasz"       "nasza"       "nasze"       "natychmiast"         "nic"       "nich"       "nie"       "niego"         "niemu"       "nim"       "nimi"       "nimi"         "obok"       "od"       "około"       "on"         "one"       "ono"       "owszem"         "pod"       "ponieważ"       "przed"       "przedtem"         "sam"       "sama"       "się"       "skąd"         "taki"       "tam"       "ten"       "to"         "tobie"       "tu"       "tutaj"       "twoi"         "was"       "wasz"       "wasza"         "was"       "waszi"       "waszystko"       "wtedy"