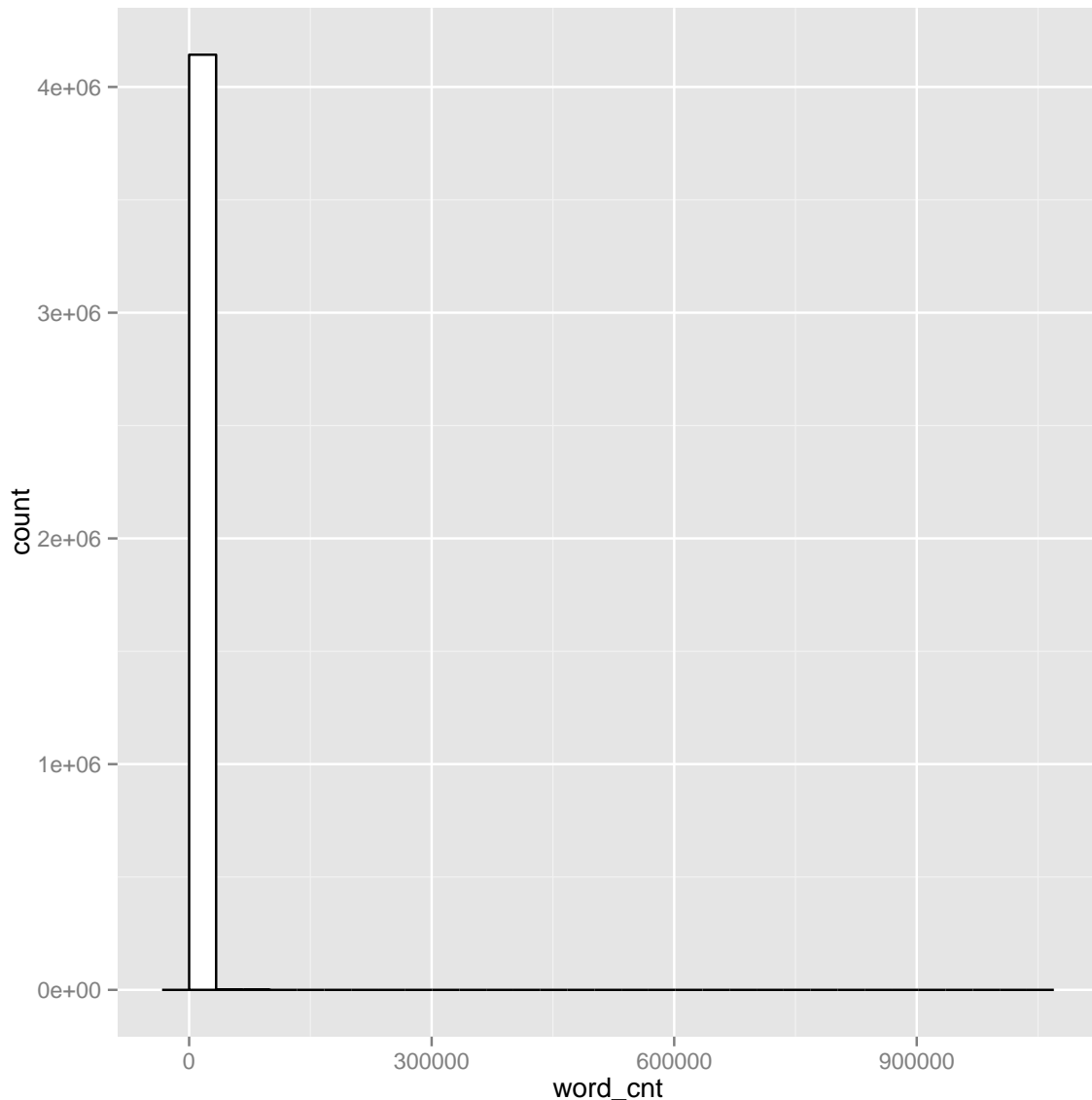


Słowa - statystyki

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] "p1516"              "jahrhundertbrücke" "111860"
[7] "kundelinga"        "bajaprogram"       "konkretniejsza"
[10] "julgneville"
```

Tabela 1: Lista dziesięciu najczęściej występujących słów.

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	5	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]	"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"
[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"	"tobą"
[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"		