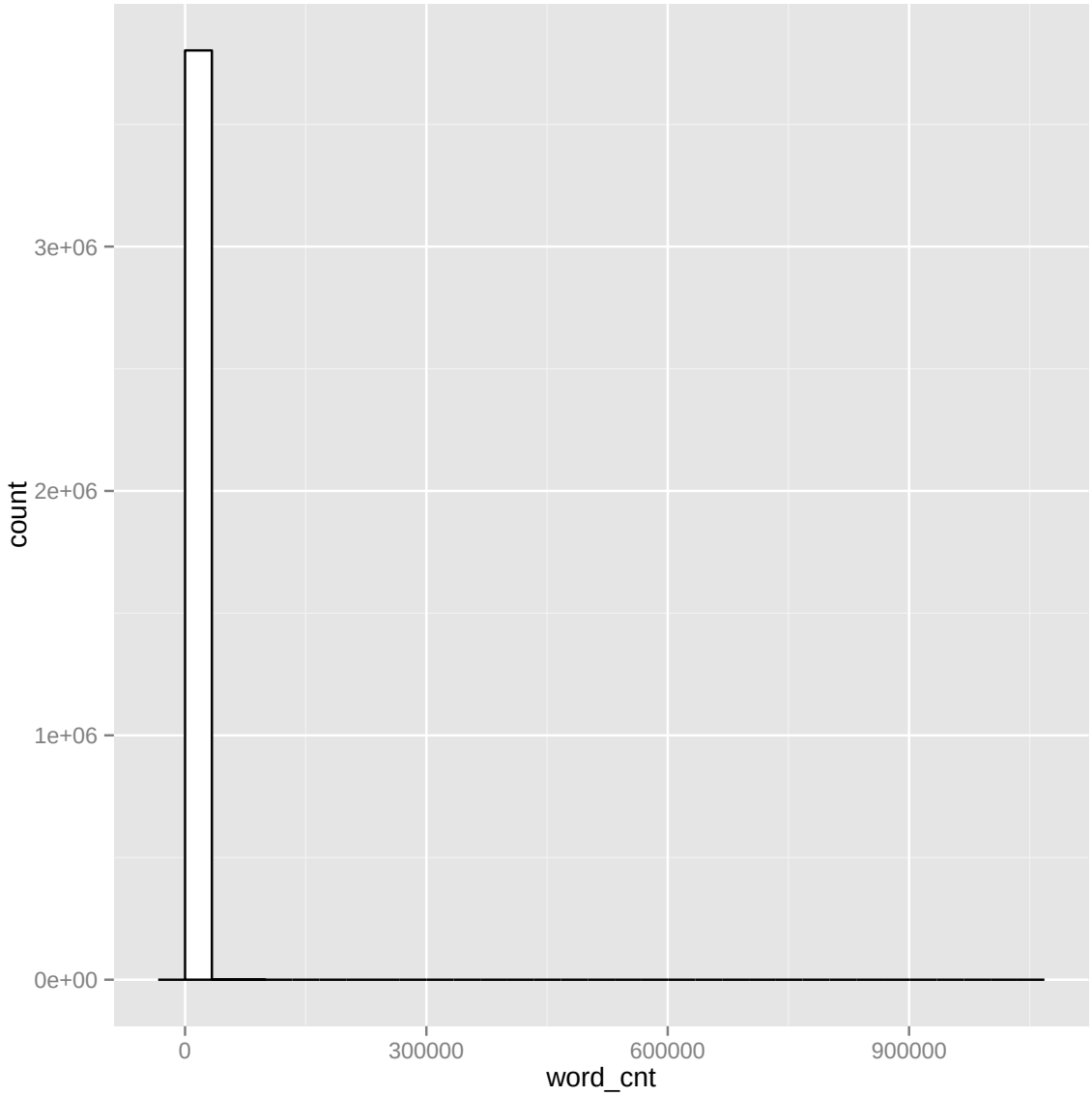
Słowa - statystyki

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"
```

```
 [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"
```