Tema nr. 3

-PSDT-

Analiza unui set de date format din 12 cărți

Partea I Incalzirea: Analiza setului de date

Partea II Tema: Determinarea cuvintelor specifice unui document/cărți pp. 21

Descrierea setului de date analizat

Setul de date cărți este format din 11 cărți, scrise in limba engleza, în care fiecare linie oferă informații cu privire la numărul versului în cadrul poeziei, versul și titlul poeziei din care face parte versul. Structura setului de date neprocesat este prezentat în Fig. 1 a).

Structura datafram-ului analizat

Setul de date: contine 12 carti

gutenberg_id	text	title	author	linie
35	The Time Machine	The Time Machine	Wells, H. G. (Herbert George)	1
35		The Time Machine	Wells, H. G. (Herbert George)	1
35	An Invention	The Time Machine	Wells, H. G. (Herbert George)	1
35	burnt brightly, and the soft radiance of the incandescent lights in the	The Time Machine	Wells, H. G. (Herbert George)	1
35	lilies of silver caught the bubbles that flashed and passed in our	The Time Machine	Wells, H. G. (Herbert George)	1
35	glasses. Our chairs, being his patents, embraced and caressed us rather	The Time Machine	Wells, H. G. (Herbert George)	1

a)

Structura datafram-ului analizat

Setul de date: contine 12 carti

gutenberg_id	title	author	linie	word
35	The Time Machine	Wells, H. G. (Herbert George)	1	the
35	The Time Machine	Wells, H. G. (Herbert George)	1	time
35	The Time Machine	Wells, H. G. (Herbert George)	1	machine
35	The Time Machine	Wells, H. G. (Herbert George)	3	an
35	The Time Machine	Wells, H. G. (Herbert George)	3	invention
35	The Time Machine	Wells, H. G. (Herbert George)	5	by

b)

Fig. 1. Structura setului de date analizat a) inainte de procesare si b) dupa procesare.

Structura setului de date după aplicarea funcției unnest_tokens() este prezentat in Fig. 1 b). Observam ca setul initial contine un numar de 144.971 de observatii, in timp ce setul de date procesat este format din 1.262.67 observatii (Fig. 2 a) si b)). Setul de date continue atat cuvinte de legatura, cat și cuvinte ce pot exprima sentimente.

```
> dim(books_df)
[1] 144971      5
```

a)

```
> dim(carti_tokenized)
[1] 1264267       5
```

b)

Fig. 2. Structura setului de date analizat a) inainte de procesare si b) dupa procesare.

Structura datafram-ului analizat

Setul de date: contine 1 carte

gutenberg_id	title	author	linie	word
35	The Time Machine	Wells, H. G. (Herbert George)	1	the
35	The Time Machine	Wells, H. G. (Herbert George)	1	time
35	The Time Machine	Wells, H. G. (Herbert George)	1	machine
35	The Time Machine	Wells, H. G. (Herbert George)	3	an
35	The Time Machine	Wells, H. G. (Herbert George)	3	invention

In Fig. 3 a) si b) este prezentată sintaxa folosită în R și rezultatul obtinut la nivelul întregului set de date.

```
> temp <- carti_tokenized %>%
+   group_by(word) %>%
+   count(word, sort = TRUE) %>%
+   distinct(word)
> dim(temp)
[1] 42159     1
```

a)

Nr.Crt.	word	n
1	the	70004
2	of	38512
3	and	29161
4	to	26867
5	in	24930

b)

Fig. 3. Prezentam a) sintaxa in R si b) rezultatul obtinut la nivelul întregului set de date.

```
temp <- carti_tokenized %>%
  group_by(word) %>%
  count(word, sort = TRUE) %>%
  head(10)
> dim(carti_tokenized)
[1] 1264267       5
```

```
> temp <- tidy_carti %>%
+   group_by(word) %>%
+   count(word, sort = TRUE) %>%
+   distinct(word)
> dim(temp)
[1] 41474     1
```

a)

Nr.Crt.	word
1	en
2	de
3	het
4	hij
5	een

b)

Fig. 3. Prezentam a) sintaxa în R și dimensiunea setului de date și b) rezultatul obtinut la nivelul întregului set de date.

Observam ca la nivelul setului de date avem un număr de 42.159 de cuvinte unice (Fig. 3 a) si b)) la nivelul setului de date. Prin urmare, ne propunem sa vedem câte cuvinte de legatura avem și care sunt cele mai folosite, respectiv câte cuvinte urmează sa analizăm.

```
> temp <- carti_cuvinte_legatura %>%
+   group_by(word) %>%
+   count(word, sort = TRUE) %>%
+   distinct(word)
> dim(temp)
[1] 685   1
```

a)

Nr.Crt.	word	n
1	the	70004
2	of	38512
3	and	29161
4	to	26867
5	in	24930
6	a	19771
7	i	18862

b)

Fig. 4. Prezentam a) sintaxa în R și dimensiunea setului de date analizat si b) rezultatul obtinut.

Astfel, avem un număr de 685 de cuvinte de legatura unice la nivelul întregului set de date (Fig. 4 a)), iar cel mai folosit cuvant de legatura este ***the*** cu 70004 apariții (Fig. 4 b)), in timp ce, numărul de cuvinte unice ce ar putea exprima sentimente este de 41474 cuvinte, iar cuvântul ***en*** este cel mai utilizat cuvant cu un numar de 7621 aparitii (Fig. 5 b)).

```
> temp <- carti_tokenized %>%
+   group_by(word) %>%
+   filter(word != '')%>%
+   count(word, sort = TRUE) %>%
+   anti_join(stop_words) %>%
+   ungroup() %>%
+   mutate(nr = row_number())
Joining with `by = join_by(word)`
> dim(temp)
[1] 41474     3
```

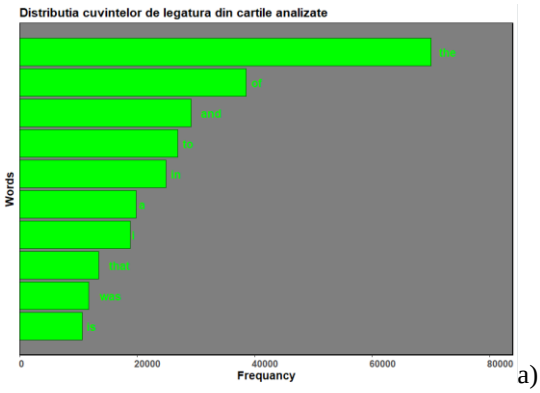
a)

Nr.Crt.	word	n	nr
1	en	7621	1
2	de	6571	2
3	het	5099	3
4	hij	4819	4
5	een	4047	5
6	zij	3917	6
7	dat	3609	7

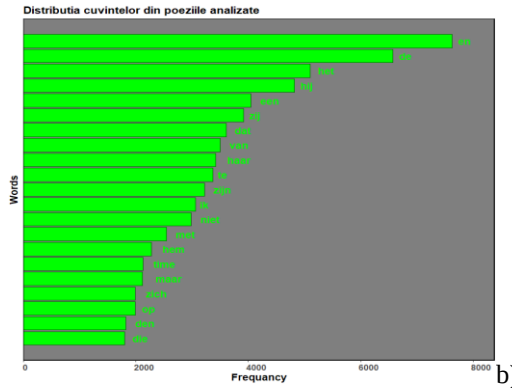
b)

Fig. 5. Prezentam a) sintaxa în R și dimensiunea setului de date analizat si b) rezultatul obtinut.

Și cum o imagine vorbește mai mult decat o mie de cuvinte în Fig. 6 a) si b) prezentăm distribuția cuvintelor de legatura și a cuvintelor ce ar putea exprima sentimente/ceva.



a)



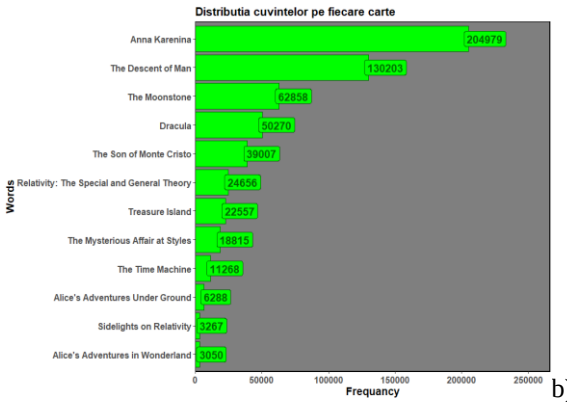
b)

Fig. 6. Distributia a) cuvintelor de legatura și b) cuvintelor ce ar putea exprima sentimente/ceva.

Analiza setului de date fără cuvinte de legatura

Nr.Crt.	title	author	n
1	Anna Karenina	Tolstoy, Leo, graf	204979
2	The Descent of Man	Darwin, Charles	130203
3	The Moonstone	Collins, Wilkie	62858
4	Dracula	Stoker, Bram	50270
5	The Son of Monte Cristo	Lermina, Jules	39007
6	Relativity: The Special and General Theory	Albert Einstein	24656
7	Treasure Island	Stevenson, Robert Louis	22557
8	The Mysterious Affair at Styles	Christie, Agatha	18815
9	The Time Machine	Wells, H. G. (Herbert George)	11268
10	Alice's Adventures Under Ground	Carroll, Lewis	6288

a)



b)

Fig. 7. Prezentam a) denumirea cărții, respectiv autorul și b) distribuția cuvintelor pe fiecare carte analizata.

In Fig. 7 b) este prezentată distribuția cuvintelor pe fiecare carte în parte. Observam ca ***Anna Karenina*** si ***The Descent of Man*** conțin cele mai multe cuvinte, respectiv 204.979 de cuvinte și 130.203 de cuvinte urmate de restul cărților, iar la polul opus se afla ***Sidelights on Relativity*** si ***Alice's Adventures in Wonderland*** cu cate 3267 de cuvinte, respectiv 3050 de cuvinte. Observa de asemenea ca apar niște cuvinte sau grupuri de litere, precum: ***en***, ***de***, ***het***, ***hij***, ***een*** etc. care nu sunt cuvinte ci mai degrabă prescurtări, notații sau un soi de cuvinte de legatura (cum ar fi de exemplu legatura van der waals care exprima forta de atractie sau respingere dintre molecule folosită cel mai probabil în cărțile științifice analizate) si pe care le vom elimina.

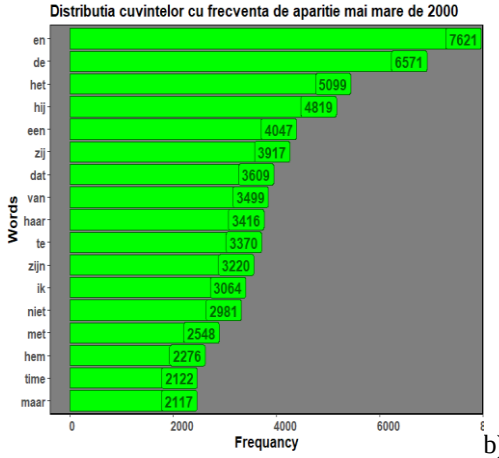
Prin urmare, vom crea o listă personalizată de cuvinte pe care o vom adauga la lista stop_words, ce va continue o serie de cuvinte sau grupuri de litere ce nu ne transmit nimic sau nu stim ce reprezinta (denim lista de cuvinte custom_stop_words (Fig. 8.)).

```
custom_stop_words <- bind_rows(data_frame(word = c('en',   'de',   'het',  'hij',  'een',  'zij',  'dat',  'van',  'haar', 'te',
                                                   'zijn', 'ik',   'niet', 'met',  'hem', 'maar', 'zich', 'op', 'den',
                                                   'aan',    'voor', 'mij', 'er', 'ook', 'er', 'als', 'naar', 'nu', 'je', 'tot', 'bij',
                                                   'dit','tm', 'uit', 'zag', 'dr', 'zou', 'gij', 'vol', 'hoe', '_k', 'heb', '_k_', 'i', "don't"),
                                    lexicon = c("custom")),
                          stop_words)
```

Fig. 8. Prezentăm sintaxa folosită în R pentru crearea listei de cuvinte personalizată.

Nr.Crt.	word	n
1	en	7621
2	de	6571
3	het	5099
4	hij	4819
5	een	4047
6	zij	3917
7	dat	3609
8	van	3499
9	haar	3416
10	te	3370

a)



b)

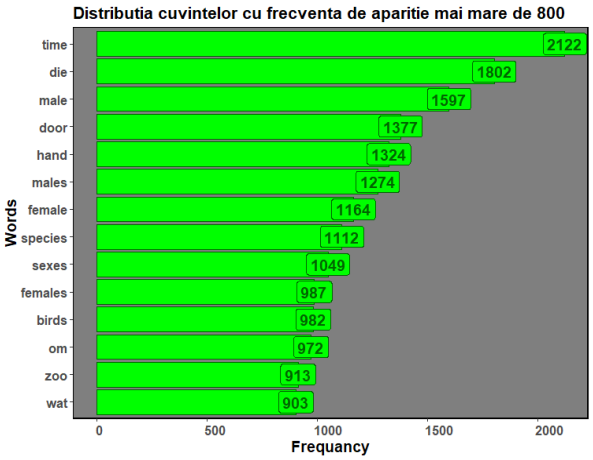
Fig. 9. Distribuția cuvintelor la nivelul cărților analizate.

Din Fig. 10 a) si b), după eliminarea cuvintelor/grupurilor de litere, observam ca cel mai utilizat cuvant este ***time*** cu un numar de 2122 aparitii, urmat de cuvântul ***die***, ***male***, ***door*** etc. De asemenea, setul de date final, după eliminarea cuvintelor/grupurilor de litere nesemnificative, conține un număr de 41.435 de cuvinte unice.

```
> tidy_carti1 <- carti_tokenized %>%
+   anti_join(custom_stop_words) %>%
+   group_by(word) %>%
+   count(word, sort = TRUE)
Joining with `by = join_by(word)`
> dim(tidy_carti1)
[1] 41435     2
```

Nr.Crt.	word	n
1	time	2122
2	die	1802
3	male	1597
4	door	1377
5	hand	1324
6	males	1274
7	female	1164
8	species	1112
9	sexes	1049

a)



b)

Fig. 10. Distribuția cuvintelor la nivelul cărților analizate după eliminarea cuvintelor/grupurilor de litere nesemnificative.

Din Fig. 11 a), după eliminarea cuvintelor/grupurilor de litere care nu transmit nimic, observam ca cel mai utilizat cuvant este ***die*** si apare in cartea ***Anna Karenina***, dar si in alte carti, urmat de o serie de cuvinte care apare in cartea ***The Descent of Man***, precum: ***male***, ***males***, ***female***, ***species*** etc.

In Fig. 11 b) sunt prezentate distribuțiile cuvintelor la nivelul fiecărei cărți în parte. Observam ca cel mai utilizat cuvant este ***time*** si apare in cartea ***The Moonstone***, dar si in alte carti, urmat de cuvantul ***die*** care apare în cartea ***Anna Karenina***, ***male*** care apare in cartea ***The Descent of Man*** etc.

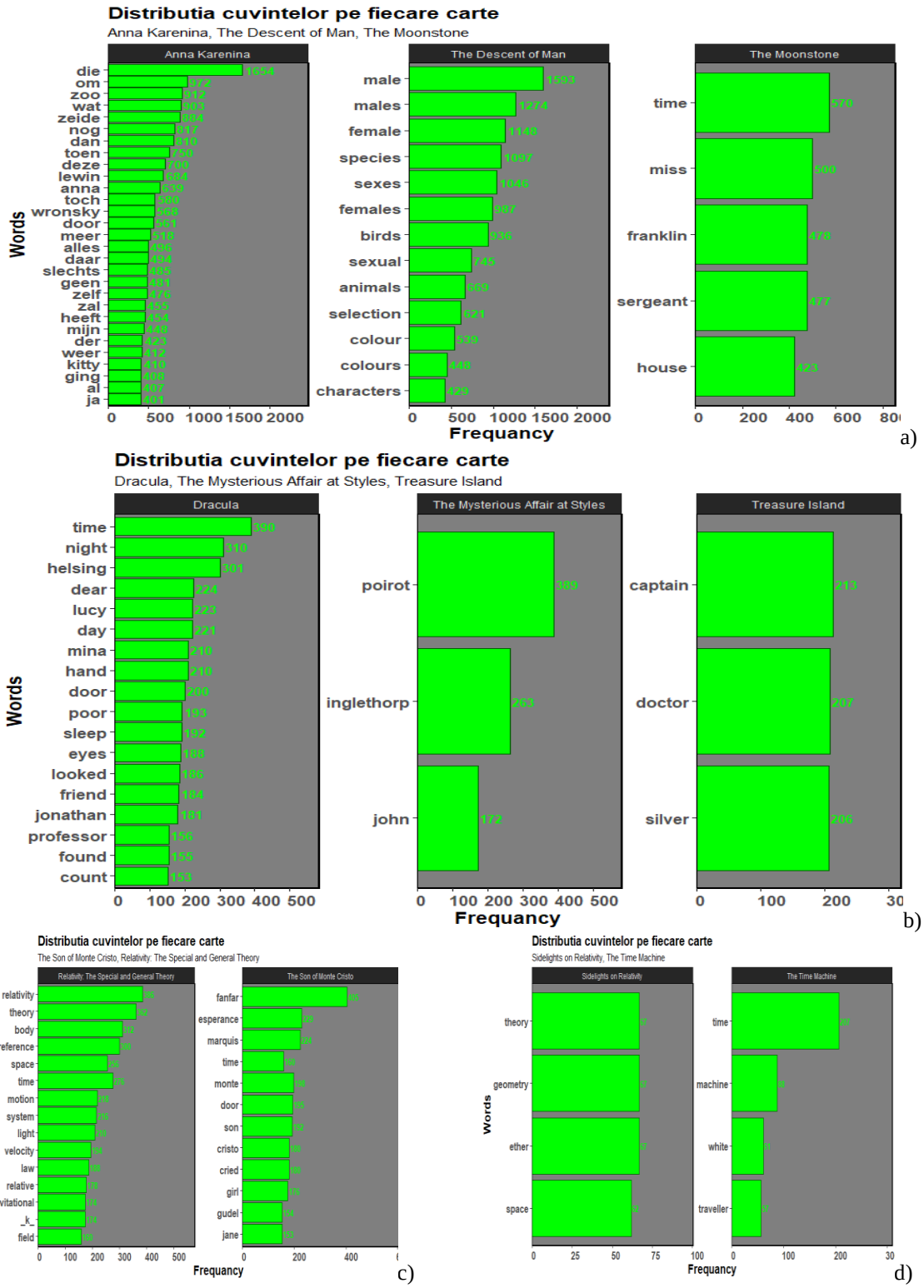
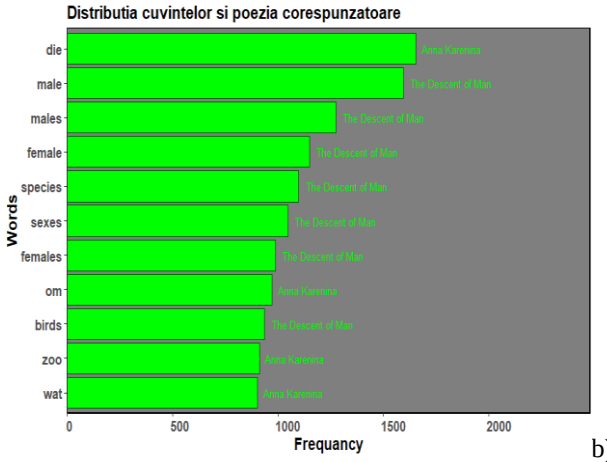


Fig. 12. Distribuția cuvintelor la nivelul cărților analizate după eliminarea cuvintelor/grupurilor de litere nesemnificative pe fiecare carte.

Nr.Crt.	title	author	word	n
1	Anna Karenina	Tolstoy, Leo, graf	die	1654
2	The Descent of Man	Darwin, Charles	male	1593
3	The Descent of Man	Darwin, Charles	males	1274
4	The Descent of Man	Darwin, Charles	female	1148
5	The Descent of Man	Darwin, Charles	species	1097
6	The Descent of Man	Darwin, Charles	sexes	1046
7	The Descent of Man	Darwin, Charles	females	987
8	Anna Karenina	Tolstoy, Leo, graf	om	972
9	The Descent of Man	Darwin, Charles	birds	936
10	Anna Karenina	Tolstoy, Leo, graf	zoo	912
11	Anna Karenina	Tolstoy, Leo, graf	wat	903
12	Anna Karenina	Tolstoy, Leo, graf	zeide	884

a)



b)

Fig. 12. Prezentam a) denumirea cărților, autorul și frecvența de apariție a cuvintelor, respectiv b) distribuția cuvintelor pe fiecare carte analizata.

In Fig. Fig. 12 a) – d) prezentăm distribuția cuvintelor care apar cu o frecvență mai mare de 50 prezente în cele douăsprezece cărți analizate: ***Anna Karenina***, ***The Descent of Man***, ***The Moonstone***, ***The Son of Monte Cristo***, ***Dracula***, ***The Mysterious Affair at Styles***, ***Relativity: The Special and General Theory***, ***Treasure Island***, ***The Time Machine***, ***Alice's Adventures Under Ground***, ***Alice's Adventures in Wonderland*** si ***Sidelights on Relativity***.

Fiecare grafic individual arată cuvintele cheie din fiecare carte și frecvența lor. De exemplu, in cartea ***Anna Karenina***: cuvântul ***die*** apare de 1654 ori, in timp ce in cartea ***The Descent of Man***: cuvintele ***male, female, species*** și ***males*** apar de 1593 ori, de 1148 ori, de 1097 ori, respective de 1274, dar apar și alte cuvinte. In cartile scrise de Albert Einstein predomina cuvinte precum: cuvintele ***relativity*** (386 de aparitii)***, space*** (256 de aparitii)***, time*** (276 de aparitii)***, referenice*** (300 de aparitii), ***theory*** (362 de aparitii) etc., iar in cartea ***Dracula***: cuvântul ***time*** apare de 390 ori, fiind, cel mai probabil, tema cartii (Fig. 13 a) si d)).

Nr.Crt.	title	author	word	n
1	The Moonstone	Collins, Wilkie	time	570
2	Dracula	Stoker, Bram	time	390
3	Relativity: The Special and General Theory	Albert Einstein	time	276
4	The Descent of Man	Darwin, Charles	time	209
5	The Time Machine	Wells, H. G. (Herbert George)	time	207
6	The Son of Monte Cristo	Lermina, Jules	time	160
7	Treasure Island	Stevenson, Robert Louis	time	133
8	The Mysterious Affair at Styles	Christie, Agatha	time	103
9	Alice's Adventures Under Ground	Carroll, Lewis	time	29
10	Sidelights on Relativity	Einstein, Albert	time	23
11	Alice's Adventures in Wonderland	Carroll, Lewis	time	22

a)

Nr.Crt.	title	author	word	n
1	Anna Karenina	Tolstoy, Leo, graf	die	1654
2	Dracula	Stoker, Bram	die	45
3	The Descent of Man	Darwin, Charles	die	44
4	The Son of Monte Cristo	Lermina, Jules	die	35
5	Treasure Island	Stevenson, Robert Louis	die	12
6	The Moonstone	Collins, Wilkie	die	7
7	The Time Machine	Wells, H. G. (Herbert George)	die	3
8	The Mysterious Affair at Styles	Christie, Agatha	die	2

b)

Nr.Crt.	title	author	word	n
1	The Descent of Man	Darwin, Charles	male	1593
2	Anna Karenina	Tolstoy, Leo, graf	male	1
3	The Moonstone	Collins, Wilkie	male	1
4	The Mysterious Affair at Styles	Christie, Agatha	male	1
5	The Time Machine	Wells, H. G. (Herbert George)	male	1

c)

Nr.Crt.	title	author	word	n
1	The Descent of Man	Darwin, Charles	female	1148
2	The Moonstone	Collins, Wilkie	female	15
3	The Time Machine	Wells, H. G. (Herbert George)	female	1

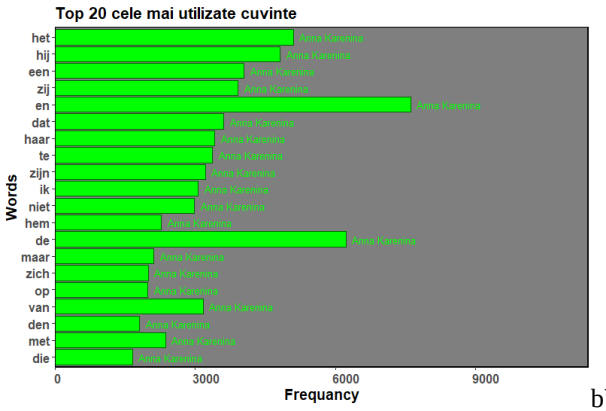
d)

Fig. 13. Frecvența de apariție a cuvintelor a) ***time***, b) ***die***, c) ***male*** si d) ***female*** prezente în cărțile analizate.

De exemplu, cuvantul ***die*** are 1654 apariții în cartea ***Anna Karenina***, în timp ce în cartea ***Dracula*** și ***The Descent of Man*** apare de 45 de ori, respective 44 de ori. Un alt cuvant ce apare în 5 cărți diferite, ***Anna Karenina***, ***The Descent of Man***, ***The Moonstone***, ***The Mysterious Affair at Styles*** is ***The Time Machine***, este ***male***. Cuvântul ***time*** apare în 11 cărți diferite, în timp ce ***die*** apare în 8 cărți diferite frecvența de apariție variind în funcție de tema abordată de autor.

Nr.Crt.	title	word	n
1	Anna Karenina	en	7614
2	Anna Karenina	de	6228
3	Anna Karenina	het	5099
4	Anna Karenina	hij	4819
5	Anna Karenina	een	4047
6	Anna Karenina	zij	3917
7	Anna Karenina	dat	3609
8	Anna Karenina	haar	3416
9	Anna Karenina	te	3370
10	Anna Karenina	zijn	3220

a)



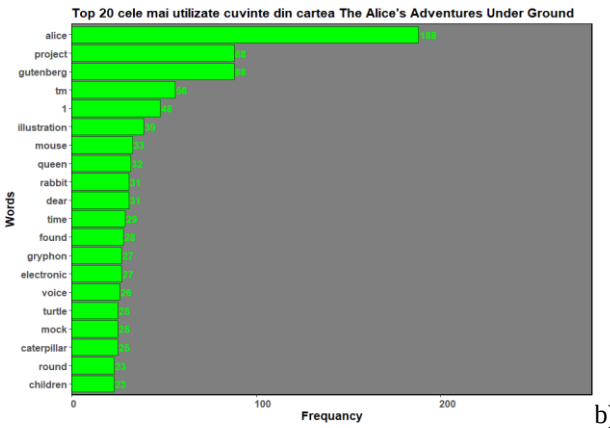
b)

Fig. 14. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Anna Karenina***.

Pentru a pune mai bine în evidenta cuvintele sau grupurile de cuvinte ce urmează a fi eliminate, in Fig. 15 – Fig. 15 a) si b) sunt prezentate distribuțiile cuvintelor pe fiecare carte în parte înainte și după eliminarea cuvintelor/grupurilor de cuvinte.

Nr.Crt.	title	word	n
1	Alice's Adventures Under Ground	alice	188
2	Alice's Adventures Under Ground	gutenberg	88
3	Alice's Adventures Under Ground	project	88
4	Alice's Adventures Under Ground	tm	56
5	Alice's Adventures Under Ground	1	48
6	Alice's Adventures Under Ground	illustration	39
7	Alice's Adventures Under Ground	mouse	33
8	Alice's Adventures Under Ground	queen	32
9	Alice's Adventures Under Ground	dear	31

a)

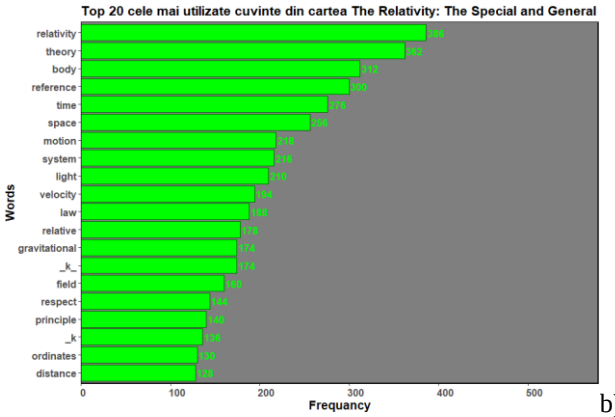


b)

Fig. 15. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Alice's Adventures Under Ground***.

Nr.Crt.	title	word	n
1	Relativity: The Special and General Theory	relativity	386
2	Relativity: The Special and General Theory	theory	362
3	Relativity: The Special and General Theory	body	312
4	Relativity: The Special and General Theory	reference	300
5	Relativity: The Special and General Theory	time	276
6	Relativity: The Special and General Theory	space	256
7	Relativity: The Special and General Theory	motion	218
8	Relativity: The Special and General Theory	system	216
9	Relativity: The Special and General Theory	light	210

a)

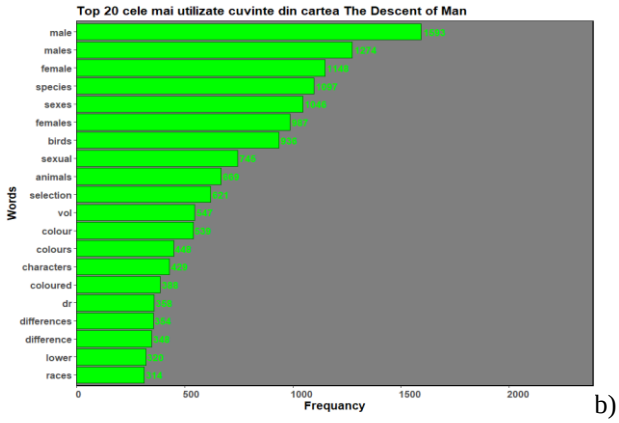


b)

Fig. 16. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Relativity: The Special and General Theory***.

Nr.Crt.	title	word	n
1	The Descent of Man	male	1593
2	The Descent of Man	males	1274
3	The Descent of Man	female	1148
4	The Descent of Man	species	1097
5	The Descent of Man	sexes	1046
6	The Descent of Man	females	987
7	The Descent of Man	birds	936
8	The Descent of Man	sexual	745
9	The Descent of Man	animals	669
10	The Descent of Man	selection	621
11	The Descent of Man	vol	547
12	The Descent of Man	colour	539

a)

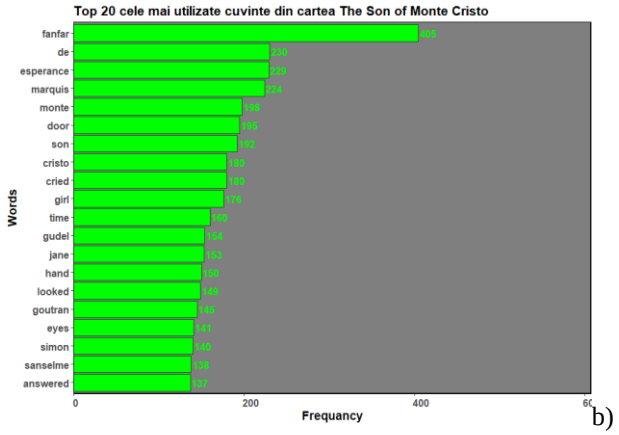


b)

Fig. 17. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***The Descent of Man***.

Nr.Crt.	title	word	n
1	The Son of Monte Cristo	fanfar	405
2	The Son of Monte Cristo	de	230
3	The Son of Monte Cristo	esperance	229
4	The Son of Monte Cristo	marquis	224
5	The Son of Monte Cristo	monte	198
6	The Son of Monte Cristo	door	195
7	The Son of Monte Cristo	son	192
8	The Son of Monte Cristo	cried	180
9	The Son of Monte Cristo	cristo	180
10	The Son of Monte Cristo	girl	176
11	The Son of Monte Cristo	time	160
12	The Son of Monte Cristo	gudel	154

a)

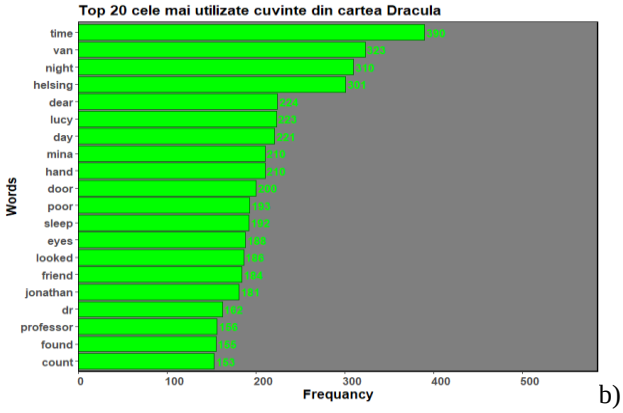


b)

Fig. 18. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***The Son of Monte Cristo***.

Nr.Crt.	title	word	n
1	Dracula	time	390
2	Dracula	van	323
3	Dracula	night	310
4	Dracula	helsing	301
5	Dracula	dear	224
6	Dracula	lucy	223
7	Dracula	day	221
8	Dracula	hand	210
9	Dracula	mina	210
10	Dracula	door	200
11	Dracula	poor	193

a)

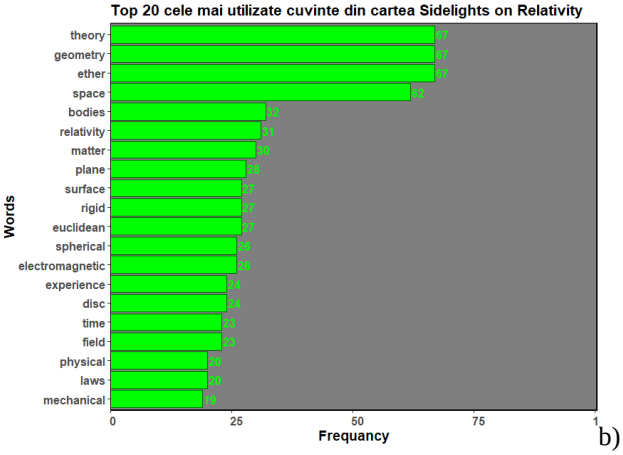


b)

Fig. 19. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Dracula***.

Nr.Crt.	title	word	n
1	Sidelights on Relativity	ether	67
2	Sidelights on Relativity	geometry	67
3	Sidelights on Relativity	theory	67
4	Sidelights on Relativity	space	62
5	Sidelights on Relativity	bodies	32
6	Sidelights on Relativity	relativity	31
7	Sidelights on Relativity	matter	30
8	Sidelights on Relativity	plane	28

a)

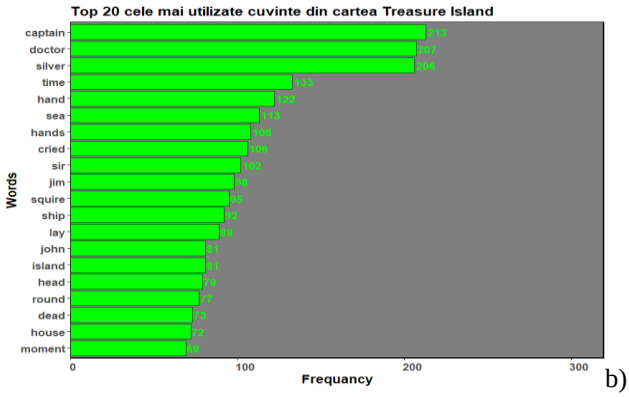


b)

Fig. 20. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Sidelights on Relativity***.

Nr.Crt.	title	word	n
1	Treasure Island	captain	213
2	Treasure Island	doctor	207
3	Treasure Island	silver	206
4	Treasure Island	time	133
5	Treasure Island	hand	122
6	Treasure Island	sea	113
7	Treasure Island	hands	108
8	Treasure Island	cried	106
9	Treasure Island	sir	102
10	Treasure Island	jim	98
11	Treasure Island	squire	95
12	Treasure Island	ship	92
13	Treasure Island	lay	89

a)

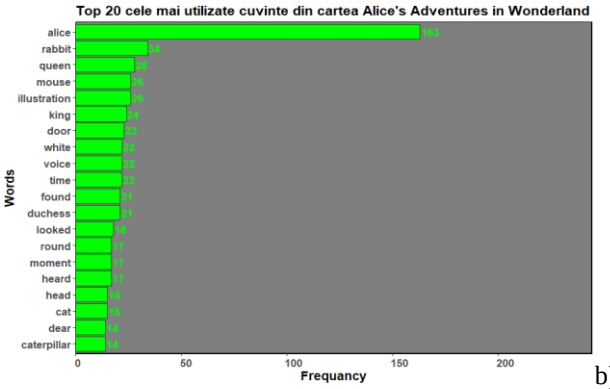


b)

Fig. 21. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Treasure Island***.

Nr.Crt.	title	word	n
1	Alice's Adventures in Wonderland	alice	163
2	Alice's Adventures in Wonderland	rabbit	34
3	Alice's Adventures in Wonderland	queen	28
4	Alice's Adventures in Wonderland	illustration	26
5	Alice's Adventures in Wonderland	mouse	26
6	Alice's Adventures in Wonderland	king	24
7	Alice's Adventures in Wonderland	door	23
8	Alice's Adventures in Wonderland	time	22
9	Alice's Adventures in Wonderland	voice	22
10	Alice's Adventures in Wonderland	white	22

a)

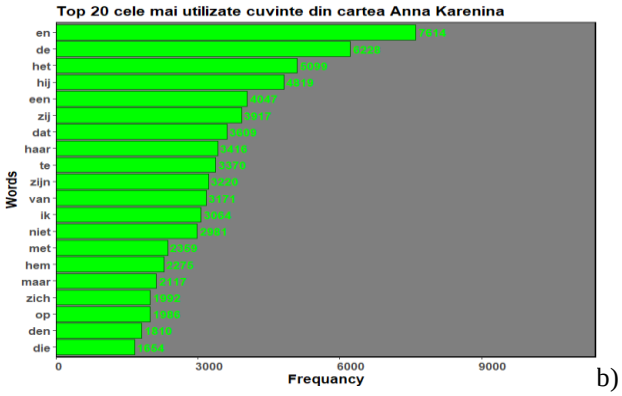


b)

Fig. 22. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Alice's Adventures in Wonderland***.

Nr.Crt.	title	word	n
1	Anna Karenina	en	7614
2	Anna Karenina	de	6228
3	Anna Karenina	het	5099
4	Anna Karenina	hij	4819
5	Anna Karenina	een	4047
6	Anna Karenina	zij	3917
7	Anna Karenina	dat	3609
8	Anna Karenina	haar	3416
9	Anna Karenina	te	3370
10	Anna Karenina	zijn	3220
11	Anna Karenina	van	3171
12	Anna Karenina	ik	3064

a)

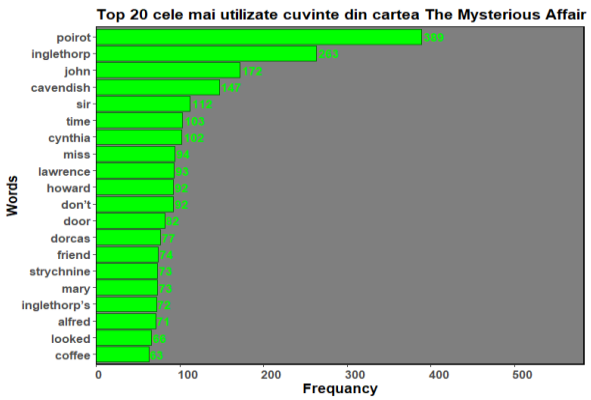


b)

Fig. 23. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***Anna Karenina***.

Nr.Crt.	title	word	n
1	The Mysterious Affair at Styles	poirot	389
2	The Mysterious Affair at Styles	inglethorp	263
3	The Mysterious Affair at Styles	john	172
4	The Mysterious Affair at Styles	cavendish	147
5	The Mysterious Affair at Styles	sir	112
6	The Mysterious Affair at Styles	time	103
7	The Mysterious Affair at Styles	cynthia	102
8	The Mysterious	miss	94

a)

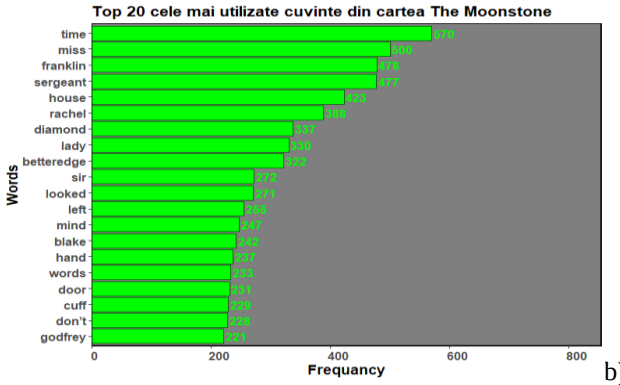


b)

Fig. 24. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***The Mysterious Affair at Styles***.

Nr.Crt.	title	word	n
1	The Moonstone	time	570
2	The Moonstone	miss	500
3	The Moonstone	franklin	478
4	The Moonstone	sergeant	477
5	The Moonstone	house	423
6	The Moonstone	rachel	388
7	The Moonstone	diamond	337
8	The Moonstone	lady	330
9	The Moonstone	betteredge	322
10	The Moonstone	sir	272
11	The Moonstone	looked	271

a)

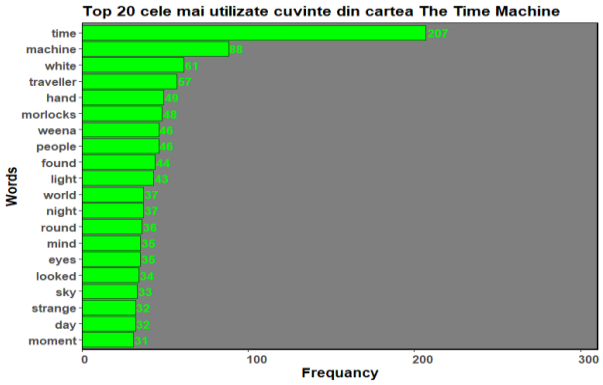


b)

Fig. 25. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***The Moonstone***.

Nr.Crt.	title	word	n
1	The Time Machine	time	207
2	The Time Machine	machine	88
3	The Time Machine	white	61
4	The Time Machine	traveller	57
5	The Time Machine	hand	49
6	The Time Machine	morlocks	48
7	The Time Machine	people	46
8	The Time Machine	weena	46
9	The Time Machine	found	44
10	The Time Machine	light	43
11	The Time Machine	night	37

a)



b)

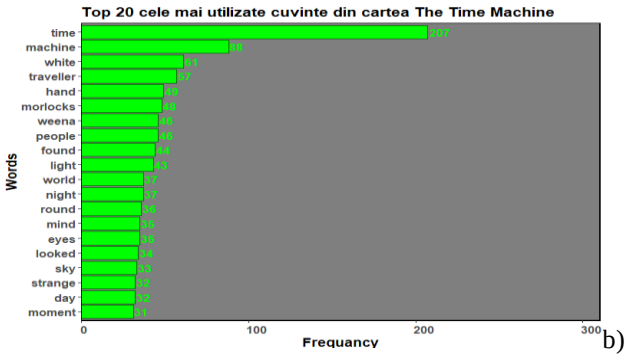
Fig. 26. a) Frecvența de apariție a cuvintelor și b) distributia acestora in ***The Time Machine***.

In Fig. 15 – Fig. 26 sunt prezentate frecvența cuvintelor, respective distribuția cuvintelor din fiecare carte analizata, unde se poate observa și cuvintele sau grupurile de litere care nu transmit nimic si care vor fi eliminate. Observam ca cele mai multe grupuri de cuvinte ce nu transmit nimic apar în ***Anna Karenina*** (en, de, het, hij etc.), ***The Descent of Man*** (dr etc.), ***The Moonstone*** (don't), ***The Son of Monte Cristo*** (de, dr etc.), ***Dracula*** (van, dr etc.), ***The Mysterious Affair at Styles*** (don't), ***Relativity: The Special and General Theory*** (_k_ etc.), ***Treasure Island*** (jim), ***The Time Machine***, ***Alice's Adventures Under Ground*** (1, tm etc.), ***Alice's Adventures in Wonderland*** si ***Sidelights on Relativity***.

In Fig. 27 – Fig. 38 a) si b) sunt prezentate frecvențele absolute a celor mai utilizate cuvintelor, respectiv distribuția cuvintelor pe fiecare carte analizata după eliminarea cuvintelor sau grupurilor de litere fără semnificație.

Nr.Crt.	title	word	n
1	The Time Machine	time	207
2	The Time Machine	machine	88
3	The Time Machine	white	61
4	The Time Machine	traveller	57
5	The Time Machine	hand	49
6	The Time Machine	morlocks	48
7	The Time Machine	people	46
8	The Time Machine	weena	46
9	The Time Machine	found	44
10	The Time Machine	light	43
11	The Time Machine	night	37
12	The Time Machine	world	37

a)

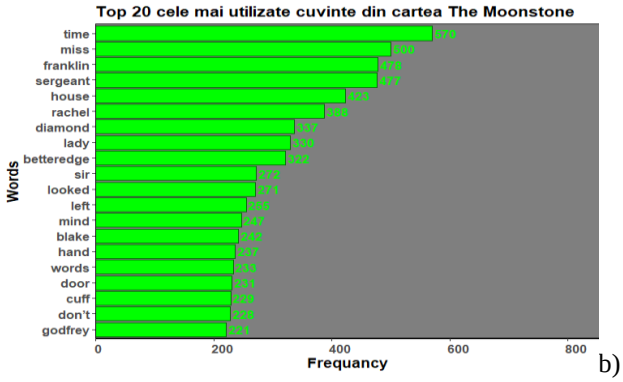


b)

Fig. 27. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***The Time Machine***.

Nr.Crt.	title	author	word	n
1	The Moonstone	Collins, Wilkie	time	570
2	The Moonstone	Collins, Wilkie	miss	500
3	The Moonstone	Collins, Wilkie	franklin	478
4	The Moonstone	Collins, Wilkie	sergeant	477
5	The Moonstone	Collins, Wilkie	house	423
6	The Moonstone	Collins, Wilkie	rachel	388
7	The Moonstone	Collins, Wilkie	diamond	337
8	The Moonstone	Collins, Wilkie	lady	330

a)

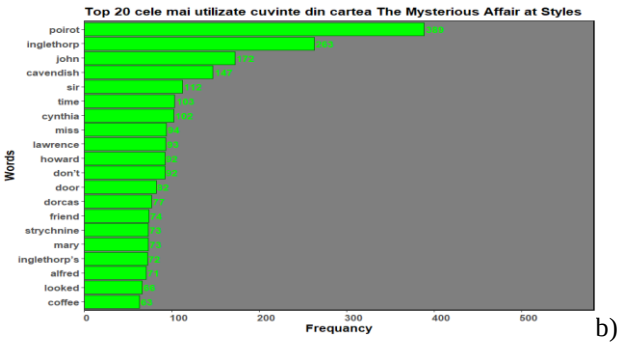


b)

Fig. 28. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***The Moonstone***.

Nr.Crt.	title	author	word	n
1	The Mysterious Affair at Styles	Christie, Agatha	poirot	389
2	The Mysterious Affair at Styles	Christie, Agatha	inglethorp	263
3	The Mysterious Affair at Styles	Christie, Agatha	john	172
4	The Mysterious Affair at Styles	Christie, Agatha	cavendish	147
5	The Mysterious Affair at Styles	Christie, Agatha	sir	112
6	The Mysterious Affair at Styles	Christie, Agatha	time	103
7	The Mysterious Affair at Styles	Christie, Agatha	cynthia	102

a)

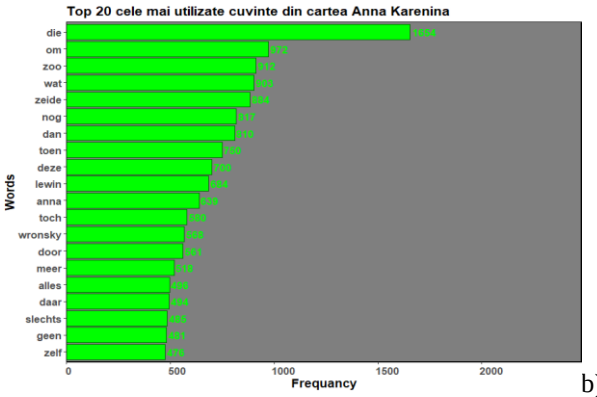


b)

Fig. 29. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***The Mysterious Affair at Styles***.

Nr.Crt.	title	author	word	n
1	Anna Karenina	Tolstoy, Leo, graf	die	1654
2	Anna Karenina	Tolstoy, Leo, graf	om	972
3	Anna Karenina	Tolstoy, Leo, graf	zoo	912
4	Anna Karenina	Tolstoy, Leo, graf	wat	903
5	Anna Karenina	Tolstoy, Leo, graf	zeide	884
6	Anna Karenina	Tolstoy, Leo, graf	nog	817
7	Anna Karenina	Tolstoy, Leo, graf	dan	810
8	Anna Karenina	Tolstoy, Leo, graf	toen	750

a)

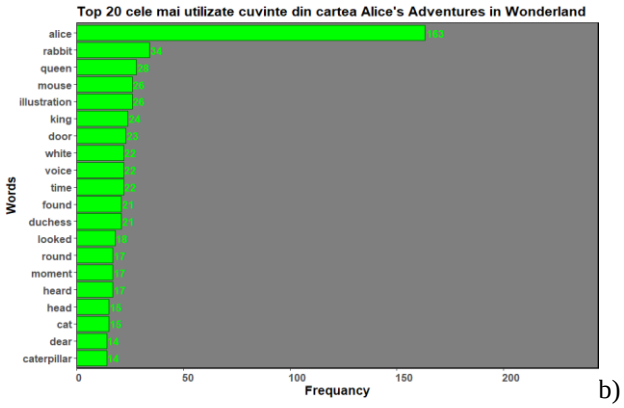


b)

Fig. 30. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Anna Karenina***.

Nr.Crt.	title	author	word	n
1	Alice's Adventures in Wonderland	Carroll, Lewis	alice	163
2	Alice's Adventures in Wonderland	Carroll, Lewis	rabbit	34
3	Alice's Adventures in Wonderland	Carroll, Lewis	queen	28
4	Alice's Adventures in Wonderland	Carroll, Lewis	illustration	26
5	Alice's Adventures in Wonderland	Carroll, Lewis	mouse	26
6	Alice's Adventures in Wonderland	Carroll, Lewis	king	24
7	Alice's Adventures in Wonderland	Carroll, Lewis	door	23
8	Alice's Adventures in Wonderland	Carroll, Lewis	time	22
9	Alice's Adventures in	Carroll, Lewis	voice	22

a)

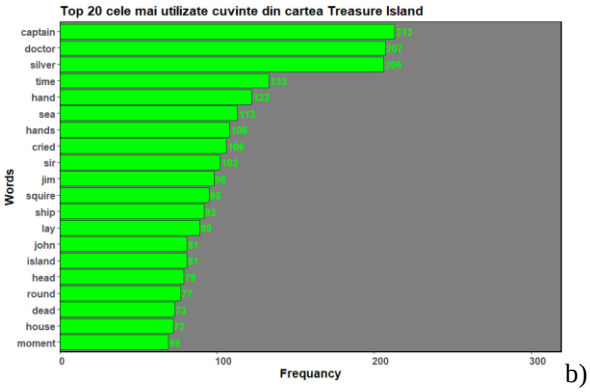


b)

Fig. 31. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Alice's Adventures in Wonderland***.

Nr.Crt.	title	author	word	n
1	Treasure Island	Stevenson, Robert Louis	captain	213
2	Treasure Island	Stevenson, Robert Louis	doctor	207
3	Treasure Island	Stevenson, Robert Louis	silver	206
4	Treasure Island	Stevenson, Robert Louis	time	133
5	Treasure Island	Stevenson, Robert Louis	hand	122
6	Treasure Island	Stevenson, Robert Louis	sea	113

a)

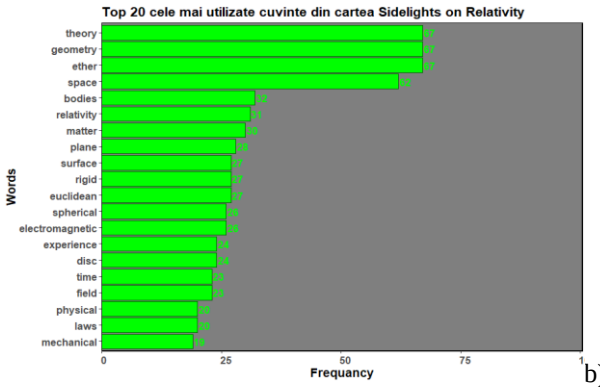


b)

Fig. 32. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Treasure Island***.

Nr.Crt.	title	author	word	n
1	Sidelights on Relativity	Einstein, Albert	ether	67
2	Sidelights on Relativity	Einstein, Albert	geometry	67
3	Sidelights on Relativity	Einstein, Albert	theory	67
4	Sidelights on Relativity	Einstein, Albert	space	62
5	Sidelights on Relativity	Einstein, Albert	bodies	32
6	Sidelights on Relativity	Einstein, Albert	relativity	31
7	Sidelights on	Einstein, Albert	matter	30

a)

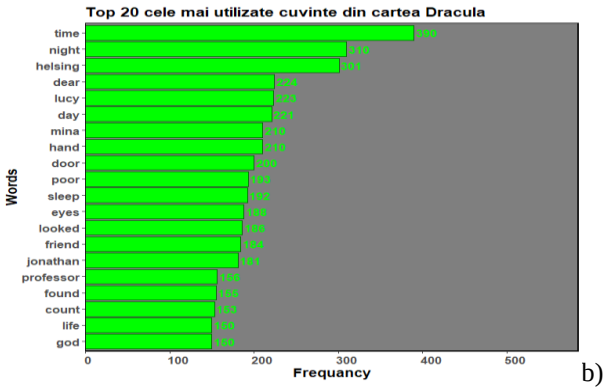


b)

Fig. 33. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Sidelights on Relativity***.

Nr.Crt.	title	author	word	n
1	Dracula	Stoker, Bram	time	390
2	Dracula	Stoker, Bram	night	310
3	Dracula	Stoker, Bram	helsing	301
4	Dracula	Stoker, Bram	dear	224
5	Dracula	Stoker, Bram	lucy	223
6	Dracula	Stoker, Bram	day	221
7	Dracula	Stoker, Bram	hand	210
8	Dracula	Stoker, Bram	mina	210
9	Dracula	Stoker, Bram	door	200

a)

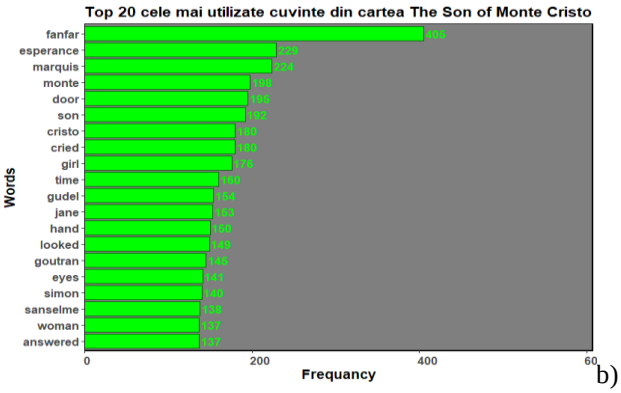


b)

Fig. 34. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Dracula***.

Nr.Crt.	title	author	word	n
1	The Son of Monte Cristo	Lermina, Jules	fanfar	405
2	The Son of Monte Cristo	Lermina, Jules	esperance	229
3	The Son of Monte Cristo	Lermina, Jules	marquis	224
4	The Son of Monte Cristo	Lermina, Jules	monte	198
5	The Son of Monte Cristo	Lermina, Jules	door	195
6	The Son of Monte Cristo	Lermina, Jules	son	192
7	The Son of	Lermina, Jules	cried	180

a)

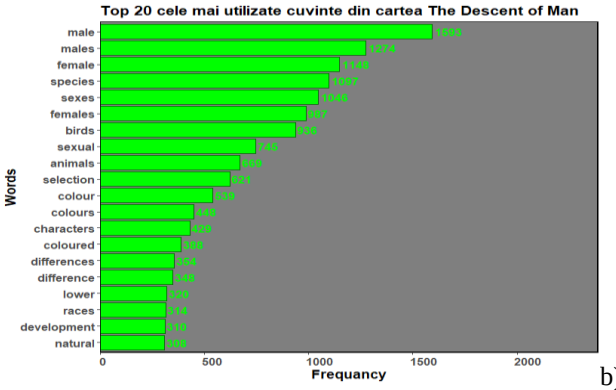


b)

Fig. 35. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***The Son of Monte Cristo***.

Nr.Crt.	title	author	word	n
1	The Descent of Man	Darwin, Charles	male	1593
2	The Descent of Man	Darwin, Charles	males	1274
3	The Descent of Man	Darwin, Charles	female	1148
4	The Descent of Man	Darwin, Charles	species	1097
5	The Descent of Man	Darwin, Charles	sexes	1046
6	The Descent of Man	Darwin, Charles	females	987
7	The Descent of Man	Darwin, Charles	birds	936
8	The Descent of Man	Darwin, Charles	sexual	745
9	The Descent of Man	Darwin, Charles	animals	669

a)

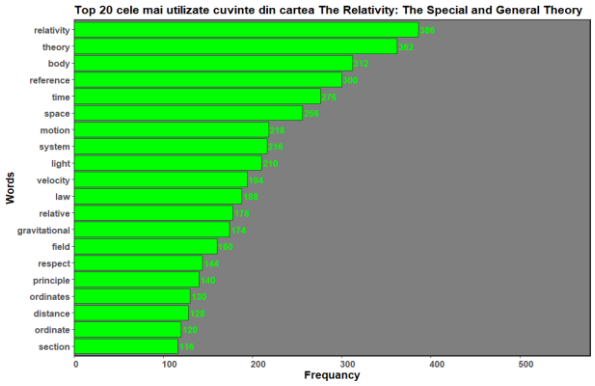


b)

Fig. 36. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***The Descent of Man***.

Nr.Crt.	title	author	word	n
1	Relativity: The Special and General Theory	Albert Einstein	relativity	386
2	Relativity: The Special and General Theory	Albert Einstein	theory	362
3	Relativity: The Special and General Theory	Albert Einstein	body	312
4	Relativity: The Special and General Theory	Albert Einstein	reference	300
5	Relativity: The Special and General Theory	Albert Einstein	time	276
6	Relativity: The Special and General Theory	Albert Einstein	space	256
7	Relativity: The Special and General Theory	Albert Einstein	motion	218
8	Relativity: The Special and	Albert Einstein	system	216

a)

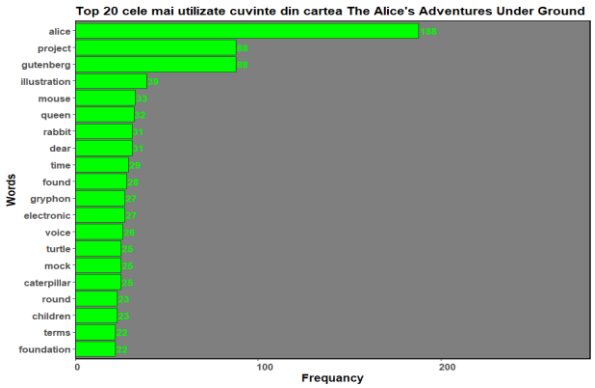


b)

Fig. 37. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Relativity: The Special and General Theory***.

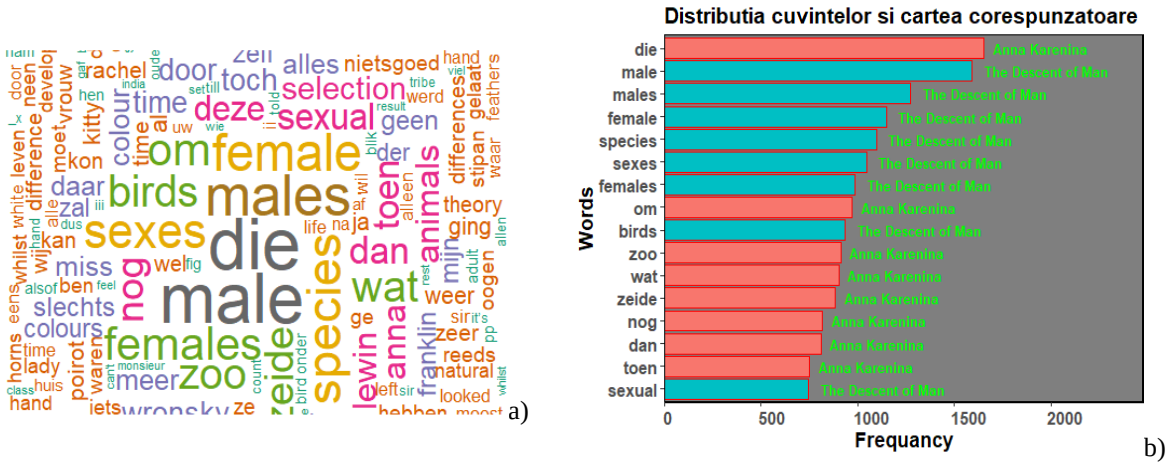
Nr.Crt.	title	author	word	n
1	Alice's Adventures Under Ground	Carroll, Lewis	alice	188
2	Alice's Adventures Under Ground	Carroll, Lewis	gutenberg	88
3	Alice's Adventures Under Ground	Carroll, Lewis	project	88
4	Alice's Adventures Under Ground	Carroll, Lewis	tm	56
5	Alice's Adventures Under Ground	Carroll, Lewis	1	48
6	Alice's Adventures Under Ground	Carroll, Lewis	illustration	39
7	Alice's Adventures Under Ground	Carroll, Lewis	mouse	33
8	Alice's Adventures Under Ground	Carroll, Lewis	queen	32
9	Alice's Adventures	Carroll, Lewis	dear	31

a)



b)

Fig. 38. a) Frecvența de apariție a cuvintelor după eliminarea cuvintelor/grupurilor de litere nesemnificative sau care nu transmit nimic si b) distributia acestora in ***Alice's Adventures Under Ground***.



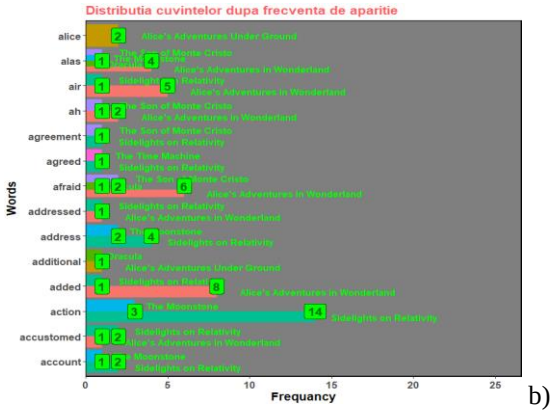
a) b)

Fig. 39. Prezentam a) norul de cuvinte și b) distribuția cuvintelor pe fiecare poezie analizata.

Pentru a ne face o imagine mai clară cu privire la cuvintele comune și frecvența lor de apariție în mai mult de o carte în Fig. 39 a) si b) sunt prezentate norul de cuvinte și frecvența de apariție a cuvintelor în diferite cărți.

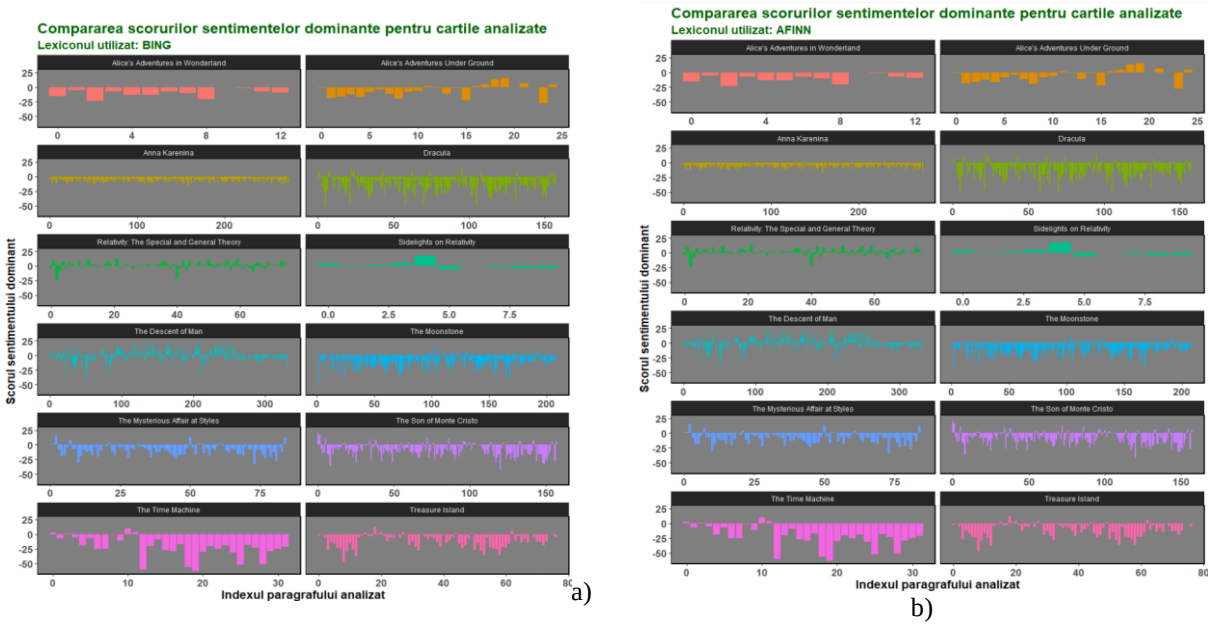
Nr.Crt.	word	title	n
1	account	Sidelights on Relativity	2
2	account	The Moonstone	1
3	accustomed	Alice's Adventures in Wonderland	1
4	accustomed	Sidelights on Relativity	2
5	action	Sidelights on Relativity	14
6	action	The Moonstone	3
7	added	Alice's Adventures in Wonderland	8
8	added	Sidelights on Relativity	1
9	additional	Alice's Adventures Under Ground	1
10	additional	Dracula	1

a)

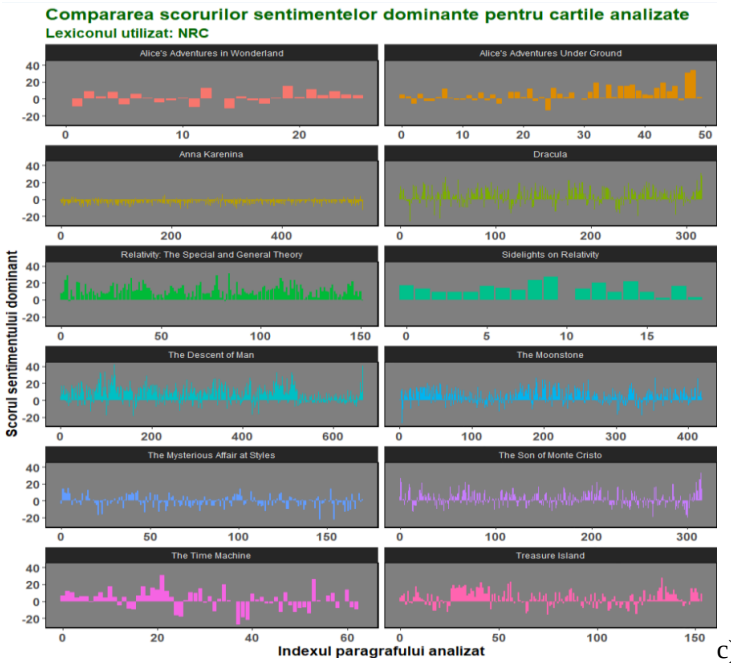


b)

Fig. 40. Prezentam a) frecvența de apariție cuvintelor în diferite cărți și b) distribuția cuvintelor comune în diferite cărți analizate.



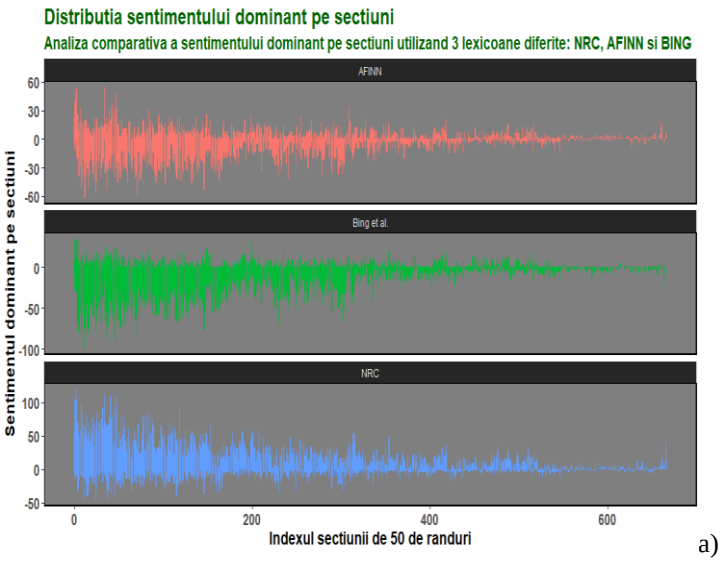
a) b)



c)

Fig. 41. Compararea scorului sentimentelor dominante pe paragrafe compuse din 50 de randuri, utilizand lexiconul a) BING, b) AFINN si c) NRC din diferite cărți analizate.

In Fig. este prezentată analiza sentimentului dominant cu privire la textul cărților analizate, utilizand trei lexicoane diferite: AFINN, BING si NRC. Observam ca, dacă folosim lexiconul NRC, sentimentul dominant, în majoritatea cărților analizate, este preponderent pozitiv, cu exceptia cartii ***Anna Karenina***. În cazul în care folosim lexicoanele AFINN si BING, sentimentul dominant, in majoritatea cartilor analizate, este negativ, cu exceptia cartilor scrise de Albert Einstein.



a)

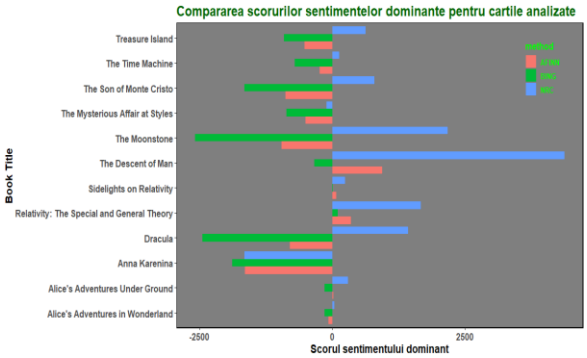
```
> get_sentiments("nrc") %>%
+   filter(sentiment %in% c("positive", "negative")) %>%
+   count(sentiment)
# A tibble: 2 x 2
  sentiment     n
  <chr>     <int>
1 negative   3316
2 positive   2308
> get_sentiments("bing") %>%
+   count(sentiment)
# A tibble: 2 x 2
  sentiment     n
  <chr>     <int>
1 negative   4781
2 positive   2005
> get_sentiments("afinn") %>%
+   mutate(sentiment = ifelse(value > 0, 'positive', 'negative')) %>%
+   count(sentiment)
# A tibble: 2 x 2
  sentiment     n
  <chr>     <int>
1 negative   1599
2 positive    878
```

b)

Fig. 42. a) Compararea scorului sentimentelor dominante pe carti, utilizand lexicone diferite: BING, AFINN si NRC si b) sintaxa utilizata in R pentru a cuantifica sentimental positive si negative pe cele trei lexicoane.

Nr.Crt.	title	author	method	sentiment
1	Alice's Adventures in Wonderland	Carroll, Lewis	AFINN	-69
2	Alice's Adventures in Wonderland	Carroll, Lewis	Bing et al.	-141
3	Alice's Adventures in Wonderland	Carroll, Lewis	NRC	44
4	Alice's Adventures Under Ground	Carroll, Lewis	AFINN	23
5	Alice's Adventures Under Ground	Carroll, Lewis	Bing et al.	-144
6	Alice's Adventures Under Ground	Carroll, Lewis	NRC	297
7	Anna Karenina	Tolstoy, Leo, graf	AFINN	-1.64e+03
8	Anna Karenina	Tolstoy, Leo, graf	Bing et al.	-1.88e+03
9	Anna Karenina	Tolstoy, Leo, graf	NRC	-1.66e+03
10	Dracula	Stoker, Bram	AFINN	-792

a)



b)

Fig. 43. Compararea scorului sentimentelor dominante pe carti, utilizand lexicon diferite: BING, AFINN si NRC.

In Fig. este prezentată analiza sentimentului dominant pe secțiuni de 50 de randuri pentru textul cărților analizate utilizând trei lexicoane diferite: AFINN, BING si NRC. Observam ca fiecare lexicon măsoară sentimentele dominante pe secțiuni de 50 de randuri în mod diferit, iar graficul arată modul în care variază sentimentul dominant pe fiecare secțiune în funcție de indexul secțiunii/paragrafului. Lexiconul AFINN acorda fiecărui cuvânt un scor numeric (cuprins intre – 5 si 5) ce indica intensitatea pozitiva sau negativa a cuvantului, în timp ce lexicoanele BING și NRC (clasifica cuvintele în 8 emoții (anger, anticipation, disgust, fear, sadness, joy, surprise, trust) și două categorii globale (pozitive și negative)) nu ia în calcul intensitatea cuvintelor ci clasifica cuvintele în doua categorii: negative is positive (Fig. 44). Observam ca sentimentul dominant al textelor analizate este negativ, daca utilizam lexicoanele AFINN si BING, respectiv pozitiv dacă folosim lexiconul NRC. Deși sentimentele paragrafelor analizate variază de-a lungul textelor analizate, observam ca sentimentele dominante, în cazul în care folosim lexicoanele AFINN și BING, este negative, ceea ce indica faptul ca autori folosesc cuvinte cu conotație negativă în text. În cazul în care folosim lexiconul NRC, observam ca fluctuatia emoțiilor în textele analizate este prezentă, dominand, de data aceasta, cuvintele cu conotatie pozitiva (sentimentele dominante fiind pozitive).

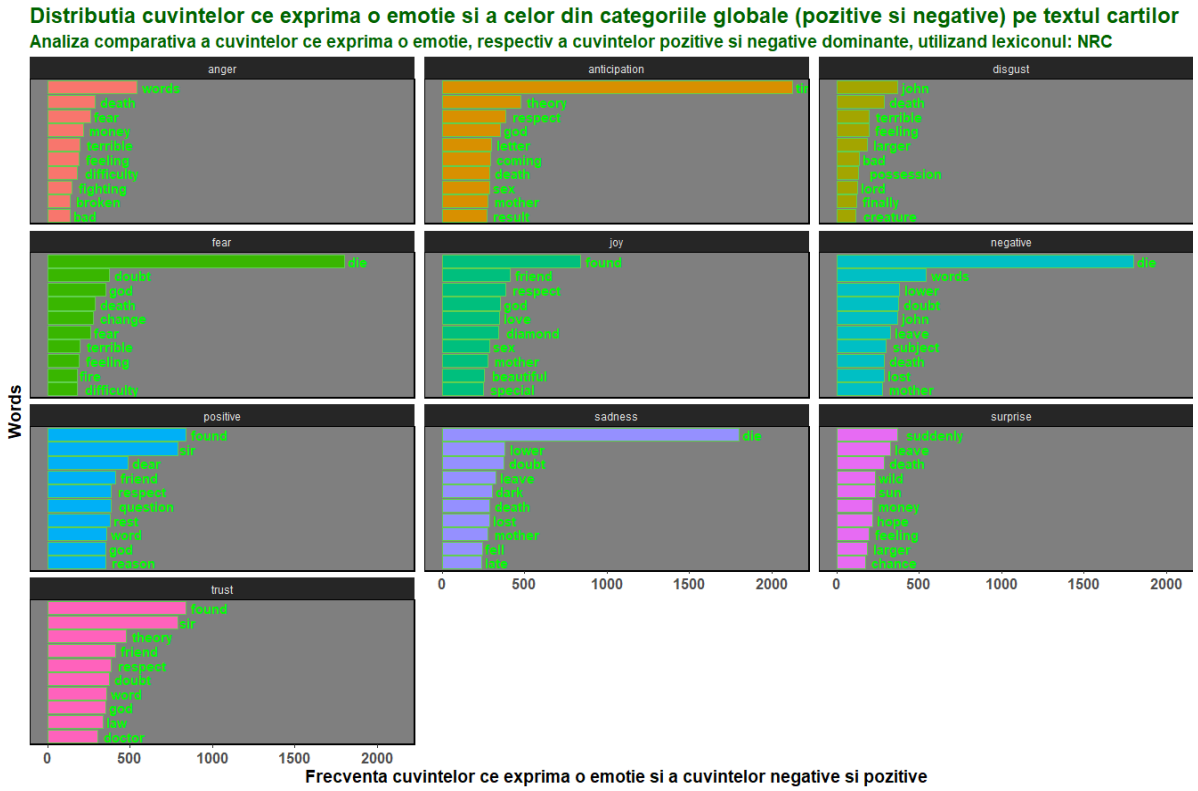


Fig. 44. Distribuția cuvintelor ce exprimă o emoție, respectiv categoriile globale (pozitive și negative) obtinute utilizand lexiconul NRC.

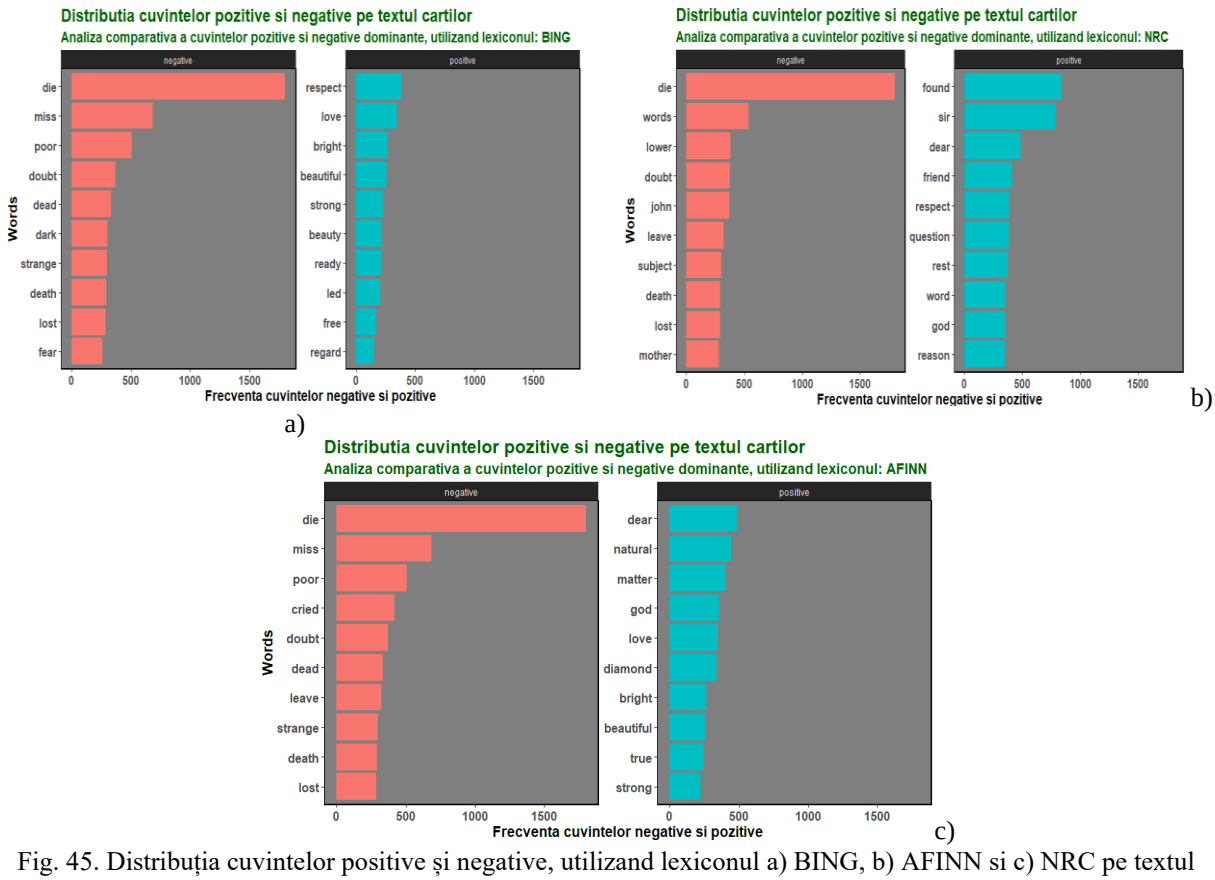


Fig. 45. Distribuția cuvintelor positive și negative, utilizand lexiconul a) BING, b) AFINN si c) NRC pe textul cărților analizate.

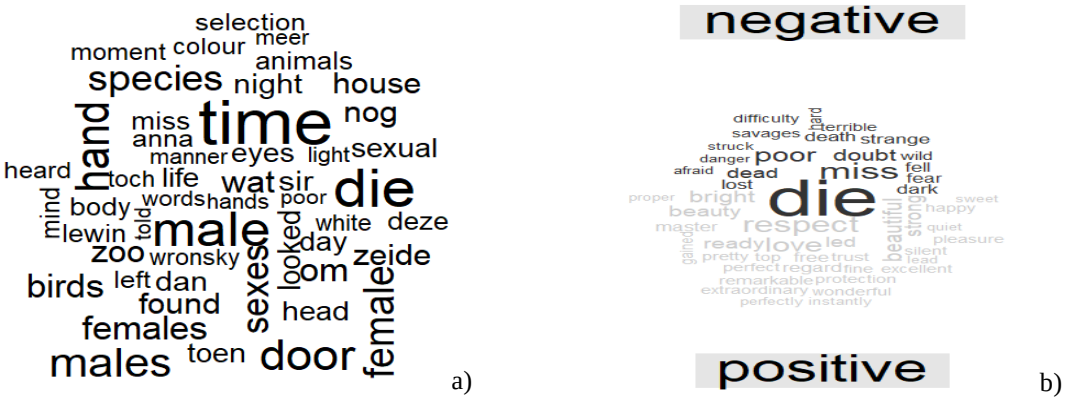


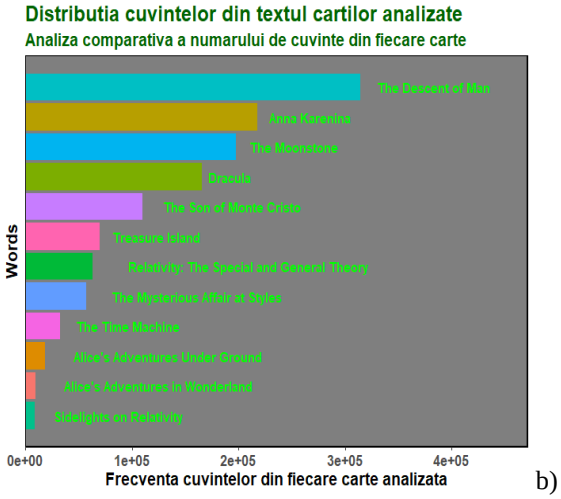
Fig. 45. Norul de cuvinte pentru textul cărților analizate.

In Fig. 46 a) si b) este prezentat norul de cuvinte pentru textul cărților analizate. Observam ca cel mai frecvent cuvant utilizat este ***time***, urmat de ***male***, ***die*** etc. în funcție de conotatia atribuita cuvintelor observăm că ***die***, ***miss***, ***poor*** etc. este un cuvânt ce face parte din categoria cuvintelor negative, in timp ce ***respect***, ***beautiful***, ***love*** etc. fac parte din categoria cuvintelor positive (Fig. 45 a) – c)).

Determinarea cuvintelor specifice unui document/cărți

Nr.Crt.	title	total
1	The Descent of Man	314222
2	Anna Karenina	217519
3	The Moonstone	197557
4	Dracula	165632
5	The Son of Monte Cristo	109596
6	Treasure Island	69640
7	Relativity: The Special and General Theory	62930
8	The Mysterious Affair at Styles	57257
9	The Time Machine	32761
10	Alice's Adventures Under Ground	18640
11	Alice's Adventures in Wonderland	9852
12	Sidelights on Relativity	8661

a)



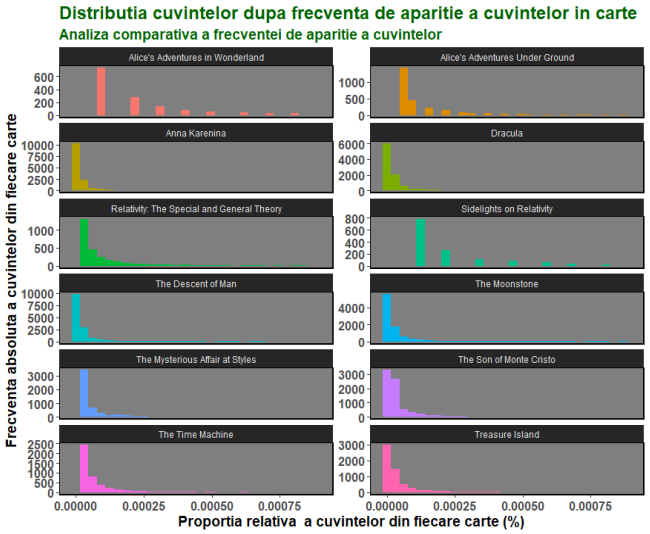
b)

Fig. 47. a) analiza comparativă a numărului de cuvinte din fiecare carte și b) distribuția cuvintelor în funcție de cartea analizata.

In Fig. 47 este prezentată distribuția cuvintelor în funcție de carte. Observam ca ***The Descent of Man*** contine cele mai multe cuvinte. Menționăm că nu s-au eliminat din setul de date analizat cuvintele de legatura.

Nr.Crt.	title	word	n	total
1	The Descent of Man	the	25682	314222
2	The Descent of Man	of	16888	314222
3	The Moonstone	the	12248	197557
4	The Descent of Man	in	8942	314222
5	Dracula	the	8101	165632
6	The Descent of Man	and	7921	314222
7	Anna Karenina	en	7614	217519
8	The Moonstone	to	6888	197557
9	The Son of Monte Cristo	the	6256	109596
10	Anna Karenina	de	6228	217519

a)

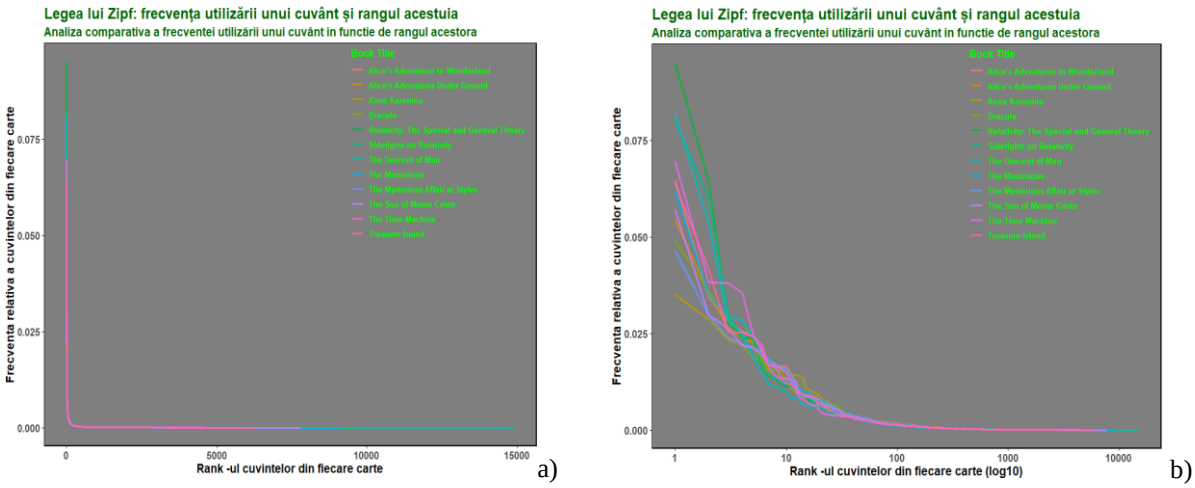


b)

Fig. 48. a) Calculul frecvenței absolute a cuvintelor și a numărului de cuvinte pe documente, respectiv b) analiza comparative a frecvenței relative a cuvintelor în textul carii analizate.

Din Fig. 48 b) se poate observa ca distribuțiile cuvintelor analizate, setul de date analizat continue atat cuvinte ce exprimă o emoție, cat și cuvinte de legatura, sunt puternic asimetrice la dreapta. Frecventa relativa a cuvintelor a fost calculata raportând numărul de cuvinte la numărul total de cuvinte din carte (Fig. 48 a)).

Pentru a obține mai multe informații despre fiecare carte (modul în care sunt distribuite cuvintele și structura textului) vom analiza frecventa relativa (df_idf) a cuvintelor în funcție de rangul asociat acestora în cadrul cărții, pentru textele analizate. Reprezentarea grafica a frecvenței relative în funcție de rangul cuvintelor pe o scala logaritmica ne poate oferii informații cu privire la respectarea sau nu a legii Zipf sau dacă se aplica sau nu legea Zipf (conform careia frecvența cuvintelor este invers proporțională cu rangul acestora, adică cele mai frecvent utilizate cuvinte tind sa fie cuvinte commune (of, the etc.), în timp ce cuvintele mai puțin utilizate sau specifice documentului sunt mai rare și au o frecventa de aparitiei mai mica).



a) b)

Fig. 49. Analiza comparative a frecvenței relative de aparitie a cuvintelor în fiecare carte a) scala utilizata este normala si b) scala utilizata este frecventa utilizata in functie de ln(rank).

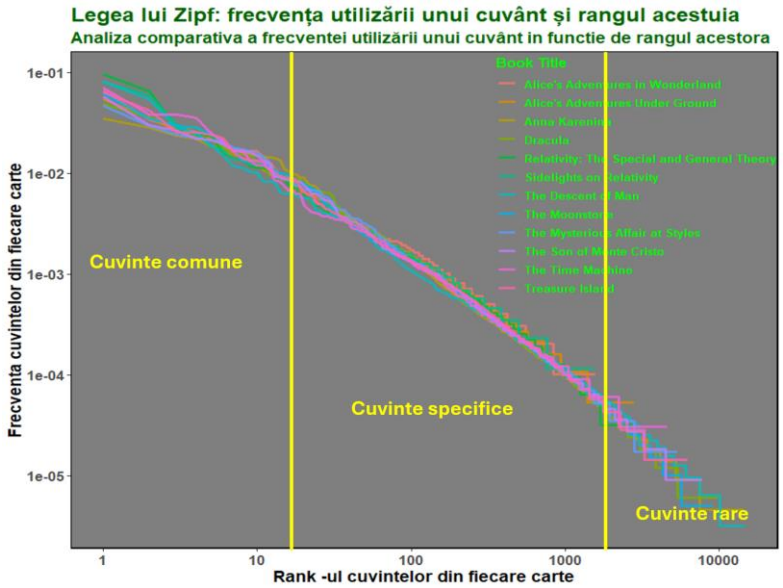
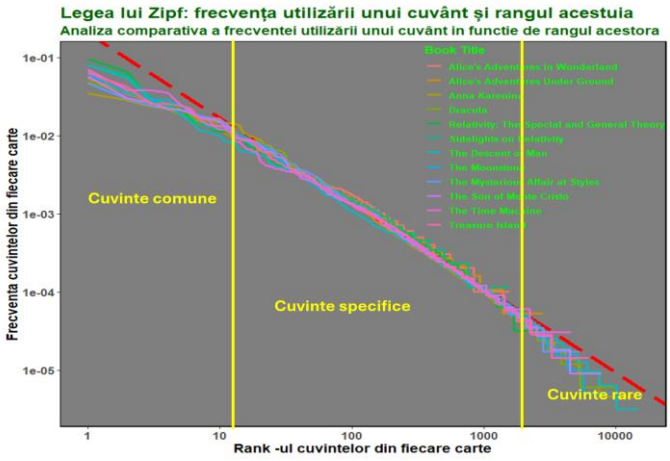


Fig. 50. Distribuția cuvintelor în funcție de valorile tf_idf.

Prin urmare, dacă se face o ordonare a cuvintelor în funcție de frecvența relativa vom observa ca aceasta scade exponential, ceea ce înseamnă ca cele mai utilizate/comune cuvintele sunt cele mai folosite, iar cuvintele rare sunt specifice cărții/textului și sunt folosite mai rar.

Nr.Crt.	title	word	n	total	rank	term frequency
1	The Descent of Man	the	25682	314222	1	0.0817
2	The Descent of Man	of	16888	314222	2	0.0537
3	The Moonstone	the	12248	197557	1	0.062
4	The Descent of Man	in	8942	314222	3	0.0285
5	Dracula	the	8101	165632	1	0.0489
6	The Descent of Man	and	7921	314222	4	0.0252
7	Anna Karenina	en	7614	217519	1	0.035
8	The Moonstone	to	6888	197557	2	0.0349
9	The Son of Monte Cristo	the	6256	109596	1	0.0571
10	Anna Karenina	de	6228	217519	2	0.0286

a)



b)

Fig. 51. Analiza comparative a a) frecvenței relative de apariție a cuvintelor în fiecare carte și b) reprezentarea grafica a frecvenței relative în funcție de rank-ul acestora.

In Fig. 51 a) si b) este prezentată o analiza comparative a frecvenței relative a cuvintelor din fiecare carte, respectiv reprezentarea grafica a acesteia în funcție de rank – ul cuvantului în cadrul fiecărei cărți pentru textele analizate. În cazul cărților analizate, observam ca legatura dintre frecventa relativa a cuvintelor și rank – ul acestora în cadrul fiecărei cărți (acordat în funcție de numărul de apariții a cuvintelor în cadrul cărții) este negativa sau inversa, putand fi

aproximata cu o legatura liniara inversa in intervalul 50 – 1000. Prin urmare, vom extrage un subset de date din eșantionul inițial pentru care rank – ul cuvintelor sa fie cuprins în intervalul (50, 1000) si vom trasa linia de regresie (Fig. 51 b)).

```
Call:
lm(formula = log10(`term frequency`) ~ log10(rank), data = rank_subset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.259623 -0.041721 -0.008573  0.031652  0.181171

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.690436   0.003514  -196.5   <2e-16 ***
log10(rank) -1.085122   0.001344  -807.4   <2e-16 ***
```

Cuvintele comune, precum ***the***, ***and*** etc., au un rang mic și o frecvență relativ mare de apariție în textele analizate, motiv pentru care apar la începutul graficului, nu sunt informative și ar trebui eliminate. Pe măsură ce ne deplasăm spre dreapta graficului, găsim cuvintele mai puțin comune, care sunt specifice fiecărui document sau cărți. Aceste cuvinte au o frecvență scăzută și un rang mare. Spre finalul graficului (dreapta), sunt plasate cuvintele rare sau unice (fie sunt cuvinte de specialitate specifice unui anumit domeniu, fie sunt cuvinte specifice subiectului dezvoltat) pentru fiecare document/carte. Prin urmare, cuvintele importante/valoroase pentru definirea subiectului/temei fiecărui document/cărți sunt cuvintele rare și specifice fiecărei cărți (si care au un scor TF-IDF mai mare, deoarece TF-IDF prioritizează cuvintele care sunt frecvente într-un document, dar apar rar în altele).

Se observă că panta liniei de regresie este mai abruptă, ceea ce sugerează că sunt câteva cuvinte folosite foarte des, în timp ce majoritatea sunt foarte rare (ceea ce este de așteptat tinand cont de faptul ca nu am eliminat cuvintele de legatura). Cu cat panta este mai abruptă cu atat textul analizat este mai simplu și conține puține cuvinte care se repeta.

Nr.Crt.	title	word	n	total	tf	idf	tf_idf
1	The Descent of Man	the	25682	314222	0.0817	0	0
2	The Descent of Man	of	16888	314222	0.0537	0	0
3	The Moonstone	the	12248	197557	0.062	0	0
4	The Descent of Man	in	8942	314222	0.0285	0	0
5	Dracula	the	8101	165632	0.0489	0	0
6	The Descent of Man	and	7921	314222	0.0252	0	0
7	Anna Karenina	en	7614	217519	0.035	1.79	0.0627
8	The Moonstone	to	6888	197557	0.0349	0	0
9	The Son of Monte Cristo	the	6256	109596	0.0571	0	0
10	Anna Karenina	de	6228	217519	0.0286	0.405	0.0116

a)

Nr.Crt.	title	word	n	tf	idf	tf_idf
1	Anna Karenina	en	7614	0.035	1.79	0.0627
2	Anna Karenina	het	5099	0.0234	2.48	0.0583
3	Anna Karenina	hij	4819	0.0222	2.48	0.0551
4	Anna Karenina	een	4047	0.0186	2.48	0.0462
5	Anna Karenina	zij	3917	0.018	2.48	0.0447
6	Anna Karenina	dat	3609	0.0166	2.48	0.0412
7	Anna Karenina	haar	3416	0.0157	2.48	0.039
8	Anna Karenina	te	3370	0.0155	2.48	0.0385
9	Anna Karenina	zijn	3220	0.0148	2.48	0.0368
10	Anna Karenina	ik	3064	0.0141	2.48	0.035

b)

Fig. 52. Analiza frecvenței relative de apariție a cuvintelor în fiecare carte (tf), idf si tf_idf a) ordonate în funcție de cuvintele cu frecvența de apariție cea mai mare și b) ordonate după valoarea tf_idf cea mai mare.

In Fig. 52 a) si b) sunt prezentate rezultatele obtinute pentru ***tf*** - frecvenței relative de apariție a cuvintelor în fiecare carte, ***idf*** = log(Numărul de documente care conțin cuvântul/Numărul total de documente – măsoară cât de rar este un cuvant in intreg setul de documente), ***tf_idf*** = tf×idf. Prin urmare, cuvintele care au o frecventa mare într-un document (tf mare) si apar in putine documente (idf mare) vor avea o valoare pentru tf_idf mare, fiind considerate cuvinte importante ce caracterizează sau sunt specifice documentului. In timp ce, cuvintele care au o frecventa mare într-un număr mare de documente, respectiv o valoare mica a idf-ului vor avea o valoare tf_idf mica, fiind considerate cuvinte comune cu o relevanță scăzută pentru diferențierea documentelor.

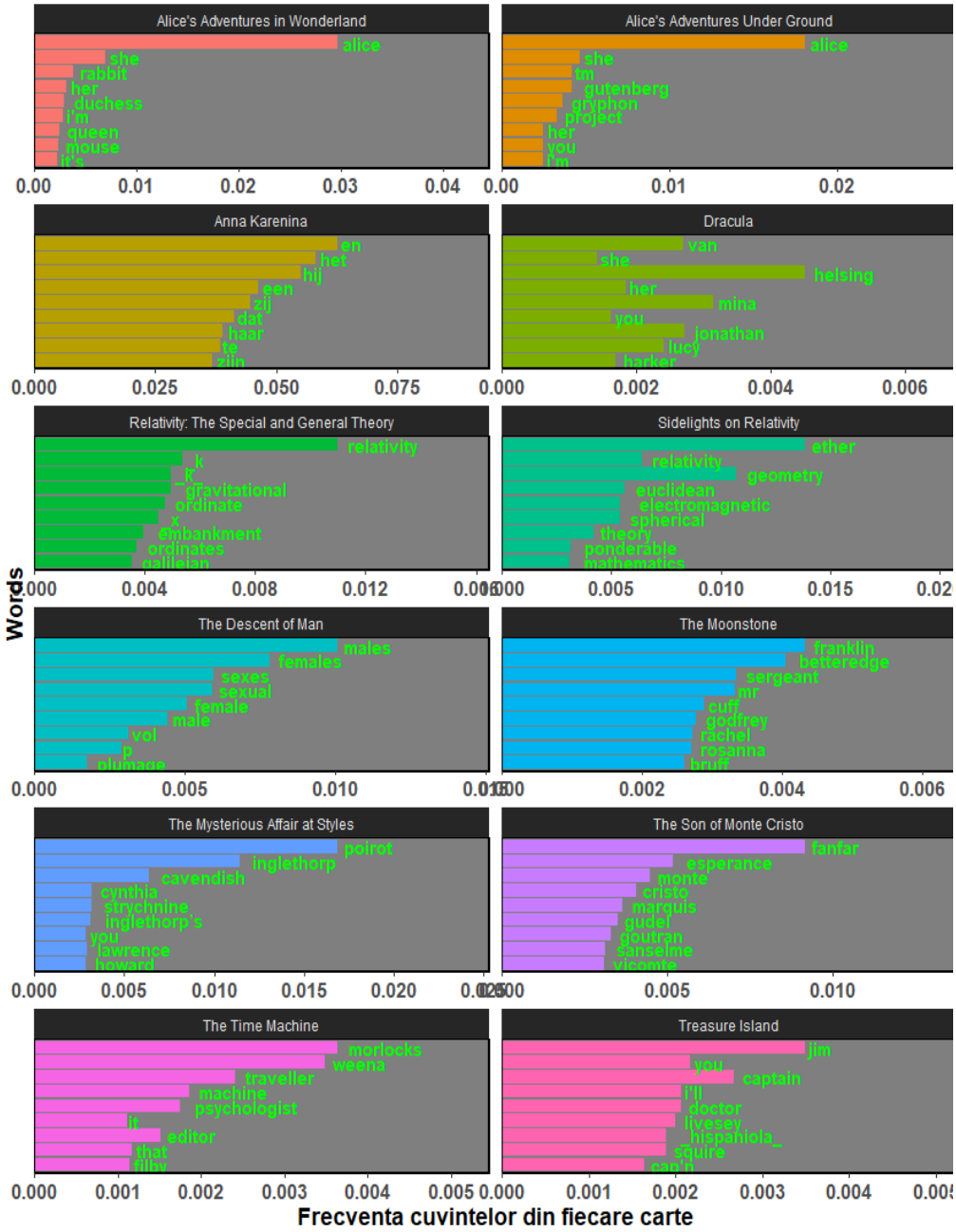


Fig. 53. Distribuția cuvintelor în funcție de valorile tf_idf.

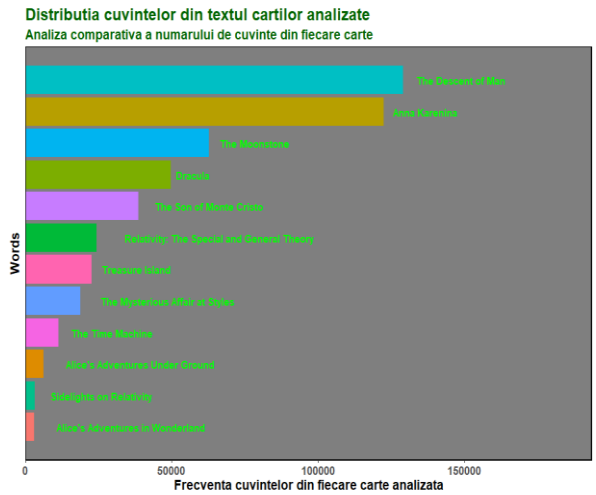
In Fig. 53 este reprezentată **frecvența celor mai comune cuvinte din fiecare carte** în funcție de valorile tf_idf (pentru texte literare și științifice). Putem observa frecvența de apariție a fiecărui **cuvânt specific într-o anumită carte**. Aceste grafice permit înțelegerea subiectului abordat în fiecare carte, deoarece cuvintele frecvențe tind să fie reprezentative pentru conținutul documentului. Astfel, cuvintele dominante in ***Alice's Adventures in Wonderland*** si ***Alice's Adventures Under Ground*** sunt nume de personaje (ex. alice, rabbit etc.) și cuvinte comune care nu sunt relevante pentru diferențierea cărților (specifice unei povesti). În cartea ***Anna Karenina*** cuvintele frecvențe sunt nume de personaje (anna, levin etc.) și cuvinte comune sau grupuri de cuvinte (en, hij, van etc.). In documente scrise de A. Einstein **(*Relativity: The Special and General Theory*** si ***Sidelights on Relativity***) observam ca cele mai frecvente cuvinte sunt termeni științifici (relativity, gravitational, mathematics, theory, electromagnetic, euclidean), dar și cuvinte fără relevanță. Aceste cuvinte indică faptul că documentul folosește cuvinte tehnice și concepte din fizică și matematică.

Analiza setului de date după ce au fost eliminate cuvintele de legatura

Astfel, prin eliminarea cuvintelor de legatura și calcularea valorilor tf_idf, cuvintele importante și reprezentative pentru fiecare carte sunt evidențiate mai bine.

Nr.Crt.	title	total
1	The Descent of Man	129006
2	Anna Karenina	122350
3	The Moonstone	62773
4	Dracula	49702
5	The Son of Monte Cristo	38645
6	Relativity: The Special and General Theory	24302
7	Treasure Island	22542
8	The Mysterious Affair at Styles	18726
9	The Time Machine	11259
10	Alice's Adventures Under Ground	6179
11	Sidelights on Relativity	3261
12	Alice's Adventures in Wonderland	3048

a)



b)

Fig. 54. a) Analiza comparativă a numărului de cuvinte din fiecare carte și b) distribuția cuvintelor în funcție de cartea analizata.

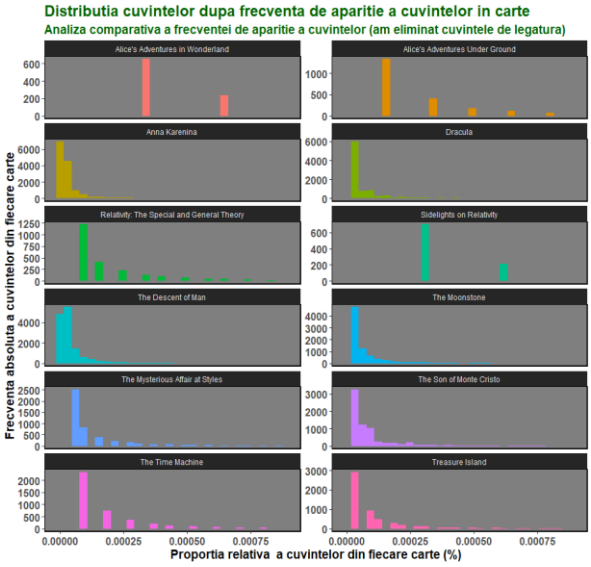
In Fig. 54 este prezentată distribuția cuvintelor în funcție de carte. Observam ca ***The Descent of Man*** contine cele mai multe cuvinte, iar la polul opus este ***Alice's Adventures in Wonderland***. Menționăm că au fost eliminate cuvintele de legatura din setul de date analizat.

Din Fig. 55 b) se poate observa ca majoritatea distributiilor cuvintelor analizate (setul de date analizat continue doar cuvinte ce exprimă o emoție sau este reprezentativ) sunt puternic

asimetrice la dreapta. Frecventa relativa a cuvintelor a fost calculata raportând numărul de cuvinte la numărul total de cuvinte din carte (Fig. 55 a)).

Nr.Crt.	title	word	n	total
1	Anna Karenina	die	1654	122350
2	The Descent of Man	male	1593	129006
3	The Descent of Man	males	1274	129006
4	The Descent of Man	female	1148	129006
5	The Descent of Man	species	1097	129006
6	The Descent of Man	sexes	1046	129006
7	The Descent of Man	females	987	129006
8	Anna Karenina	om	972	122350
9	The Descent of Man	birds	936	129006
10	Anna Karenina	zoo	912	122350

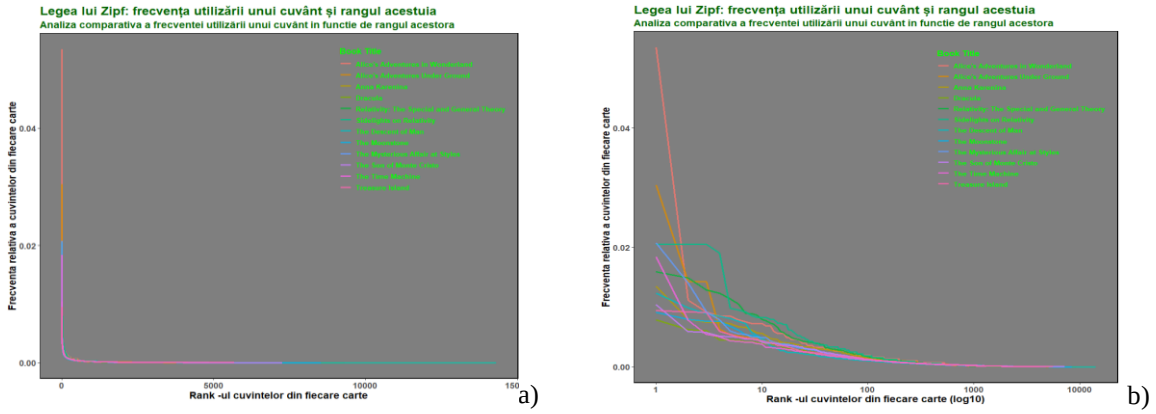
a)



b)

Fig. 55. a) Calculul frecvenței absolute și a numărului total de cuvinte pe fiecare document și b) distribuția cuvintelor după frecventa relativa de aparitie.

In Fig. 57 – Fig. 58 a) si b) sunt prezentate frecvențele relative (df_idf) în funcție de rangul cuvintelor pe o scala logaritmica. Observam ca frecvența cuvintelor este invers proporțională cu rangul acestora, adică cele mai frecvent utilizate cuvinte tind sa fie cuvinte specifice documentelor (die, male, female, spicies etc.), în timp ce cuvintele mai rare au o frecvență de apariție mult mai mica. Prin urmare, dacă se face o ordonare a cuvintelor în funcție de frecvența relativa (df_idf) vom observa ca aceasta scade exponențial și urmează o lege de tip putere, ceea ce înseamnă ca cuvintele specifice sunt cele mai folosite, iar cuvintele rare sunt specifice cărții/textului și sunt folosite mai rar.



a) b)

Fig. 56. Analiza comparative a frecvenței relative de aparitie a cuvintelor în fiecare carte a) scala utilizata este normala si b) scala utilizata este frecventa utilizata in functie de ln(rank).

Din Fig. 57 – Fig. 58 b) se observa ca are loc o scadere exponențială a frecvenței relative a cuvintelor în raport cu rangul acestora pentru toate cartele analizate. Cu alte cuvinte, pe masura ce rangul cuvantului crește (adică cuvantul are o frecventa mai mica de aparitie, este

mai rar), frecvența de apariție scade exponențial sau semnificativ fiind în concordanță cu legea lui Zipf.

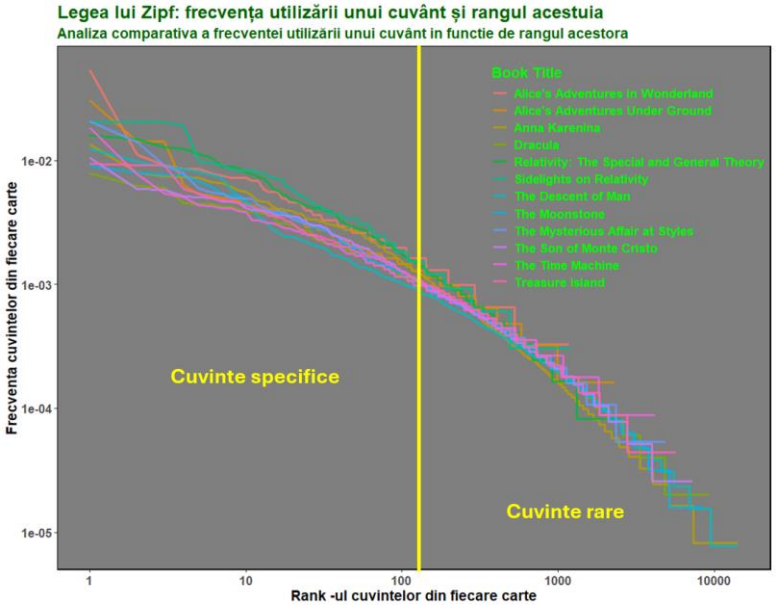
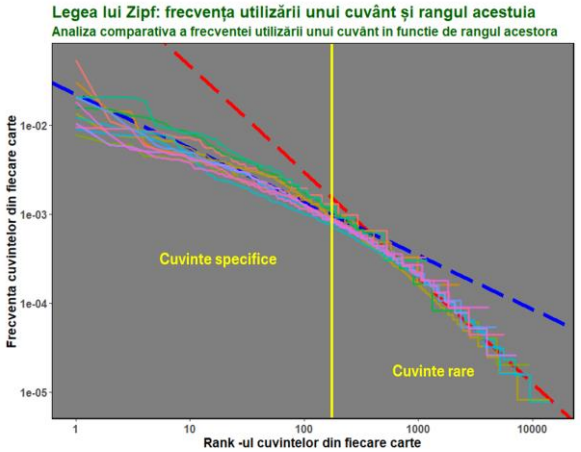


Fig. 57. Frecventa relativa (df_idf) a cuvintelor în funcție de rangul acestora.

Prin urmare, distribuția cuvintelor, după eliminarea cuvintelor de legatura sau commune, urmează o legea de distribuție de tip Zipf, confirmand faptul că cuvintele rare au o frecvență de apariție mica comparativ cu cele specifice care au o frecvență de apariție ridicată.

Nr.Crt.	title	word	n	total	rank	term frequency
1	Anna Karenina	die	1654	122350	1	0.0135
2	The Descent of Man	male	1593	129006	1	0.0123
3	The Descent of Man	males	1274	129006	2	0.00988
4	The Descent of Man	female	1148	129006	3	0.0089
5	The Descent of Man	species	1097	129006	4	0.0085
6	The Descent of Man	sexes	1046	129006	5	0.00811
7	The Descent of Man	females	987	129006	6	0.00765
8	Anna Karenina	om	972	122350	2	0.00794
9	The Descent of Man	birds	936	129006	7	0.00726
10	Anna Karenina	zoo	912	122350	3	0.00745

a)



b)

Fig. 58. Analiza comparativa a a) frecvenței relative de apariție a cuvintelor în fiecare carte și b) reprezentarea grafica a frecvenței relative în funcție de rank-ul acestora.

In Fig. 59 a) si b) sunt prezentate rezultatele obtinute pentru ***tf*** - frecvenței relative de apariție a cuvintelor în fiecare carte, ***idf*** = log(Numărul de documente care conțin cuvântul/Numărul total de documente – măsoară cât de rar este un cuvant in intreg setul de documente), ***tf_idf*** = tf×idf. Observam ca după eliminarea cuvintelor de legatura apar cuvintele specifice cărților analizate și cuvintele rare. Ordonarea descrescătoare după valorile tf_idf arată

cuvintele specifice fiecărui document și importanța lor în raport cu restul colecției de documente, respectiv modul în care fiecare cuvânt contribuie la diferențierea documentelor în funcție de conținutul lor. Cuvintele selectate ca fiind cele mai relevante pentru fiecare document după eliminarea cuvintelor comune și a celor care nu contribuie la diferențierea documentelor sunt prezentate în coloana word.

Nr.Crt.	title	word	n	total	tf	idf	tf_idf
1	Anna Karenina	die	1654	122350	0.0135	0.405	0.00548
2	The Descent of Man	male	1593	129006	0.0123	0.875	0.0108
3	The Descent of Man	males	1274	129006	0.00988	2.48	0.0245
4	The Descent of Man	female	1148	129006	0.0089	1.39	0.0123
5	The Descent of Man	species	1097	129006	0.0085	0.405	0.00345
6	The Descent of Man	sexes	1046	129006	0.00811	1.79	0.0145
7	The Descent of Man	females	987	129006	0.00765	2.48	0.019
8	Anna Karenina	om	972	122350	0.00794	2.48	0.0197
9	The Descent of Man	birds	936	129006	0.00726	0.288	0.00209
10	Anna Karenina	zoo	912	122350	0.00745	1.79	0.0134

a)

Nr.Crt.	title	word	n	tf	idf	tf_idf
1	Alice's Adventures in Wonderland	alice	163	0.0535	1.79	0.0958
2	Alice's Adventures Under Ground	alice	188	0.0304	1.79	0.0545
3	The Mysterious Affair at Styles	poirot	389	0.0208	2.48	0.0516
4	Sidelights on Relativity	ether	67	0.0205	1.79	0.0368
5	The Mysterious Affair at Styles	inglethorp	263	0.014	2.48	0.0349
6	Sidelights on Relativity	geometry	67	0.0205	1.39	0.0285
7	Relativity: The Special and General Theory	relativity	386	0.0159	1.79	0.0285
8	The Son of Monte Cristo	fanfar	405	0.0105	2.48	0.026
9	The Descent of Man	males	1274	0.00988	2.48	0.0245
10	Anna Karenina	om	972	0.00794	2.48	0.0197

b)

Fig. 59. Analiza frecvenței relative de apariție a cuvintelor în fiecare carte (tf), idf si tf_idf a) ordonate în funcție de cuvintele cu frecvența de apariție cea mai mare și b) ordonate după valoarea tf_idf cea mai mare.

Spre exemplu, cuvântul *die* în ***Anna Karenina*** are o frecvență de 0.0135, ceea ce înseamnă că reprezintă aproximativ 1.35% din totalul cuvintelor din acel document. Raritatea unui cuvânt în colecția de documente este data de indicatorul idf. Cu cât un cuvânt este mai rar în alte documente, cu atât scorul idf este mai mare. De exemplu, cuvântul *males* din ***The Descent of Man*** are un idf de 2.48, ceea ce indică faptul că este relativ rar în colecția de documente si specific pentru aceasta carte/document. Importanța relativă a unui cuvânt într-un document este data de df_idf care tine cont atât de frecvența sa în acel document, cât și de raritatea sa în întreaga colecție. Valorile mari ale acestui indicator sunt utile pentru diferențierea documentul în colecția de documente. Spre exemplu, cuvantul *male* are o valoare a tf_idf de 0.0245, ceea ce arată că este un cuvânt frecvent în acest document (tf este 0.00988) și este rar în alte documente (idf este 2.48). Sugerând că *male* este specific și important pentru *The Descent of Man*, indicând tema abordata de autor (Fig. 59).

Curba observata (Fig. 57) poate fi împărțită în doua porțiuni liniare pentru care putem trasa linia de regresie (genera un model de regresie). Astfel, am construit doua modele de regresie pentru cele doua porțiuni (prima portiune a curbei cu valori ale rangului cuprins intre (0, 150) și a doua porțiune a curbei cu valori cuprinse între (150, 10000). Prima portiune a curbei are o panta mai putin abrupta (lina albastra de regresie), sugerand ca frecvența cuvintelor scade mai lent odată cu creșterea rangului, respective vocabularul folosit este mai diversificat. A doua porțiune a curbei are o panta mult mai abrupta (linia de regresie rosie), sugerand o

scădere rapidă a frecvenței relative odată cu creșterea rangului, respective un vocabular mai specific.

```
> rank_subset <- freq_by_rank %>%
+   filter(rank < 150, rank > 10)
> lm(log10(`term frequency`) ~ log10(rank), data = rank_subset) %>%
+   coef()
(Intercept) log10(rank)
 -1.6521907  -0.6045483
```

```
> rank_subset <- freq_by_rank %>%
+   filter(rank < 10000, rank > 150)
> lm(log10(`term frequency`) ~ log10(rank), data = rank_subset) %>%
+   coef()
(Intercept) log10(rank)
 -0.1426184  -1.1889740
```

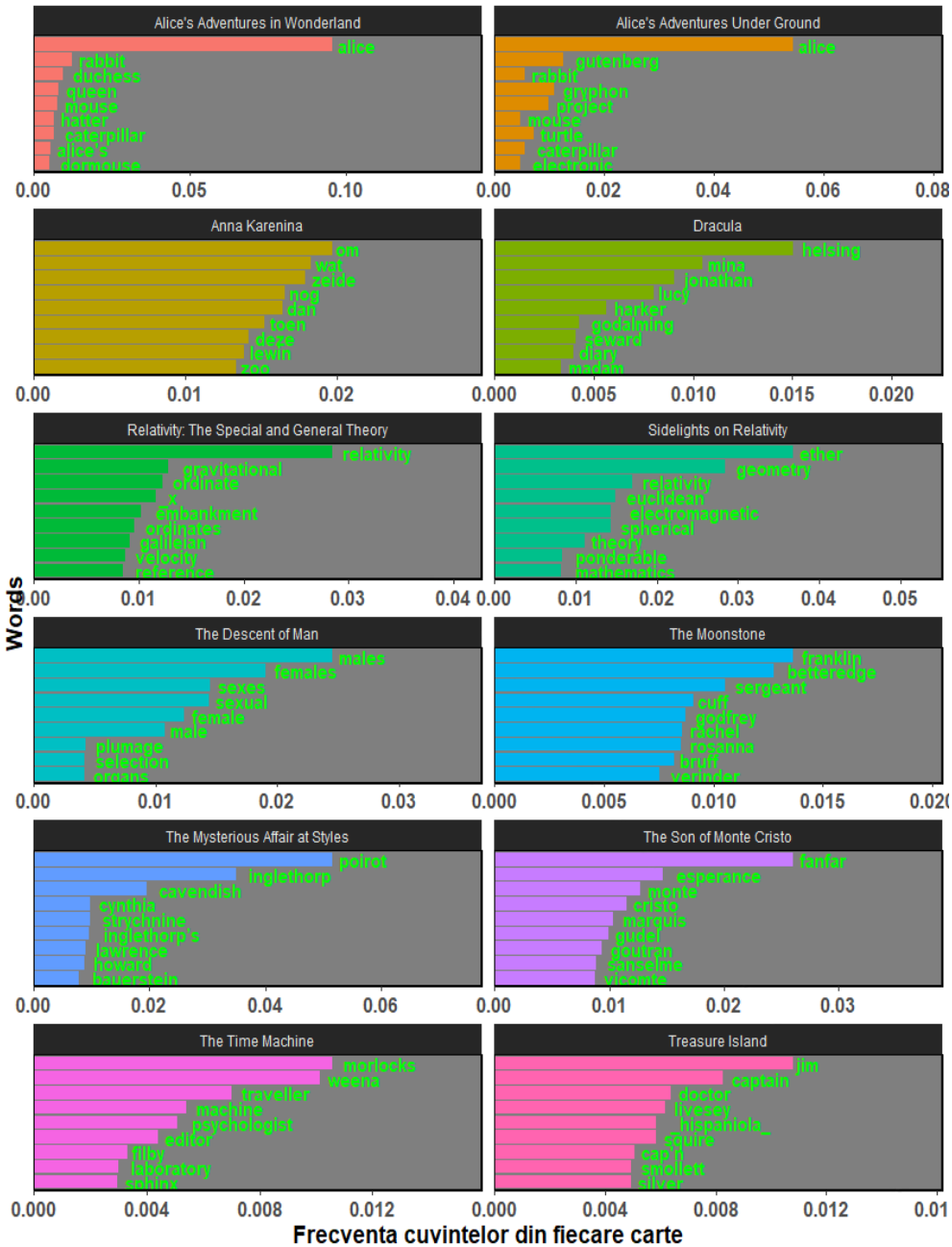


Fig. 60. Distribuția cuvintelor în funcție de valorile tf_idf.

După eliminarea **cuvintele de legătură**, în setul de date vom avea doar cuvintele care adaugă informații relevante și utile pentru înțelegerea și diferențierea documentelor, adică **cuvintele mai semnificative** din cărțile analizate. De exemplu, intr-un document stiintific, cum sunt cartile scrise de A. Einstein, după eliminarea cuvintelor de legătură/comune, cuvinte precum: *gravitation, space, time* etc. vor avea o pondere mai mare în analiza frecvenței și a valorilor df_idf, oferind o descriere mai clară a subiectului. Într-un document de literatură, cum ar fi ***Dracula***, după eliminarea cuvintelor comune, vor rămâne cuvinte precum *Jonathan, Harker*, etc., reflectând mai bine subiectul cărții și personajele (Fig. 60).