Tema nr. 5

-PSDT-

Analiza a 4 carti/documente diferite scrise in limba engleza

Descrierea setului de date analizat

Scopul temei este de a analiza 4 topicuri/subiecte si asocierea dintre acestea cu conținutul detaliat al textului corespunzator celor 4 carti. Cartile analizate sunt: ***Dracula***, ***Frankenstein***, ***The Picture of Dorian Gray*** si ***The Call of the Wild***. In cadrul temei s-a analizat probabilitatea ca un cuvant sa apartina unui subiect (beta), respect care sunt cele mai importante cuvinte pentru un topic specific. De asemenea, s-a analizat probabilitatea ca un capitol/carte sa apartina unui anumit topic, asocierea dintre subiecte si capitolele/cartile analizate. Indicatorul gamma va fi utilizat pentru a identifica subiectele dominante din capitole/carti, in timp ce beta permite identificarea subiectelor principale si a vocabularului utilizat de autor.

Setul de date analizat în cadrul temei a fost realizat prin descarcarea (de pe platforma https://www.gutenberg.org/ebooks/search/) si liprea textului/continutului a 4 carti scrise de diferiti autori, in limba engleza si transformarea setului de date intr-un corpus cu ajutorul pachetul ***quanteda***. Corpusul ***books_corpus*** obtinut a fost analizat folosind functii specifice pachetului ***quanteda***, cat si functii specifice altor pachete.

```
> books_df
# A tibble: 196,877 × 5
   gutenberg_id text                                                    title author linie
          <dbl> <chr>                                                   <chr> <chr>  <int>
 1           35 "I am going to press the lever, and off the machine will… The … Wells…     1
 2           35 "vanish, pass into future Time, and disappear. Have a go… The … Wells…     2
 3           35 "thing. Look at the table too, and satisfy yourselves th… The … Wells…     3
 4           35 "trickery. I don't want to waste this model, and then be… The … Wells…     4
 5           35 "quack.""                                                The … Wells…     5
 6           35 ""                                                       The … Wells…     6
 7           35 "There was a minute's pause perhaps. The Psychologist se… The … Wells…     7
 8           35 "speak to me, but changed his mind. Then the Time Travel… The … Wells…     8
 9           35 "his finger towards the lever. "No," he said suddenly. "… The … Wells…     9
10           35 "hand." And turning to the Psychologist, he took that in… The … Wells…    10
# i 196,867 more rows
```

Funcția din pachetul topicmodels care implementează algoritmul LDA permite extragerea de subiecte/teme dintr-un set de date de tip text. Rezultatul acestei functii ofera informatii cu privire la matricile asociate subiectelor si documentelor, distributia cuvintelor in functie de subiect, respectiv distributia temelor in functie de document (in cazul nostru fiecare carte reprezinta un document/fieare capitol din carte reprezinta un document).

Nr.Crt.	topic	term	beta
1	1	buck	3.05e-193
2	2	buck	0.00407
3	3	buck	6.49e-207
4	4	buck	2.59e-202
5	1	thornton	4.77e-195
6	2	thornton	0.00297
7	3	thornton	7.41e-41
8	4	thornton	6.52e-203
9	1	dorian	6.7e-196
10	2	dorian	0.0025

a)

Nr.Crt.	topic	term	beta
1	1	day	0.00467
2	1	die	0.00386
3	1	sea	0.00376
4	1	found	0.0034
5	1	mountains	0.00289
6	1	ice	0.0028
7	1	september	0.00261
8	1	fire	0.00257
9	1	father	0.00257
10	1	return	0.0025

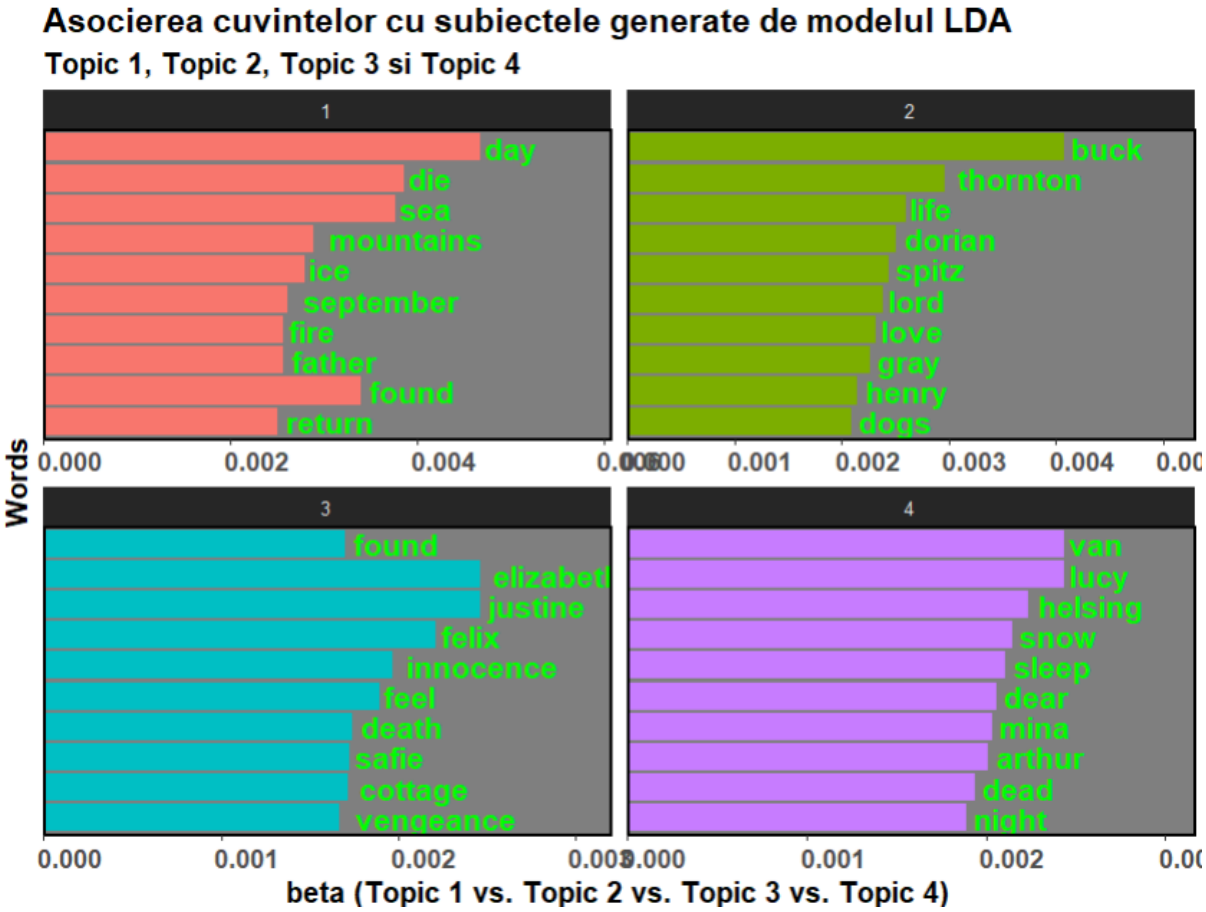
b)

Fig. 1. Asocierea cartilor analizate cu topicurile create prin modelul LDA a) valoarea indicatorului beta pentru fiecare carte și b) reprezentarea grafica a distributiei cartilor dupa indicatorul gamma in functie de topicul abordat.

In Fig. 1 sunt prezentati termeni asociați cu un anumit topic identificat cu ajutorul unui model de **LDA** (Latent Dirichlet Allocation). Fiecare termen apare împreună cu valoarea indicatorului **beta**, care reprezintă probabilitatea ca termenul respectiv să fie asociat cu subiectul analizat.

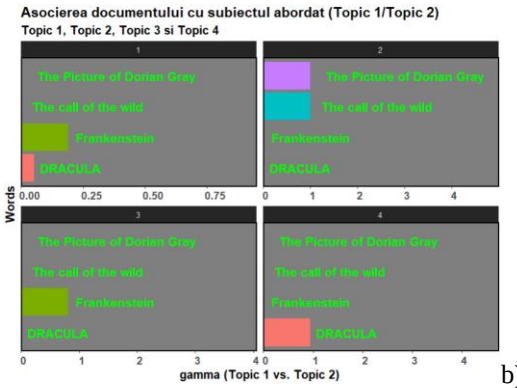
In Fig. 1 este reprezentarea grafica, **top 10 termeni** cei mai reprezentativi pentru fiecare dintre cele **4 topicuri** analizate identificate prin modelul LDA, in functie de valorile **beta**. Principalele cuvinte care apar in **Topicul 1,** sunt: *day, die, sea, mountains, ice, september, fire, father, found, return etc*. Acest subiect pare să se focalizeze pe **aventură**, **natură** și **emoții dramatice**. Spre deosebire de Topicul 1, in **Topicul 2** termeni principali, sunt: *buck, thornton, life, dorian, spitz, lord, love, gray, henry, dogs*, ceea ce indica faptul ca subiectul pare să fie asociat cu o poveste despre animale și relații. Termeni precum *buck* și *thornton* provin din cartea *The Call of the Wild*, care este o poveste despre câini și natură. Termeni ca *life, love* și *lord* ar putea sugera teme morale sau filozofice.

Topicul 3 prezinta urmatori termeni principali: *found, elizabeth, justine, felix, innocence, feel, death, safie, cottage, vengeance.* Acest subiect include elemente dramatice, posibil legate de povestea *Frankenstein*. Termeni precum *elizabeth, justine, felix* sunt nume de personaje, iar *innocence, death, vengeance* indică teme de moralitate, tragedie și răzbunare. In timp ce, in **Topicul 4** termeni principali sunt: *van, lucy, helsing, snow, sleep, dear, mina, arthur, dead, night*, care sugereaza ca subiectul este asociat cu povestea *Dracula*. Termeni precum *van (Van Helsing), lucy, mina, arthur* sunt personaje centrale, iar *dead, night, sleep* reflectă teme gotice, de groază sau vampirism.



Nr.Crt.	document	topic	gamma
1	The call of the wild	1	4.48e-06
2	The Picture of Dorian Gray	1	3.25e-06
3	DRACULA	1	0.0523
4	Frankenstein	1	0.19
5	The call of the wild	2	1
6	The Picture of Dorian Gray	2	0.992
7	DRACULA	2	2.03e-06
8	Frankenstein	2	3.31e-06
9	The call of the wild	3	4.48e-06
10	The Picture of Dorian Gray	3	0.00798

a)



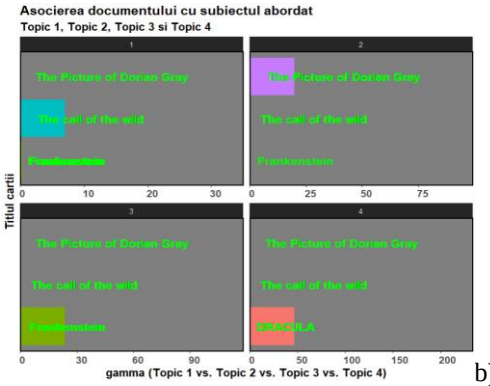
b)

Fig. 2. Asocierea cartilor analizate cu topicurile create prin modelul LDA a) valoarea indicatorului gamma pentru fiecare carte și b) reprezentarea grafica a distributiei cartilor dupa indicatorul gamma in functie de topicul abordat.

In Fig. 2 este reprezentata asocierea **fiecărui document/carte** cu **subiectele principale (Topic 1, 2, 3 și 4)**, in functie de valorile gamma, care indică proporția fiecărui subiect într-un document/carte. Observam ca documentele Dracula si Frankenstein, au o asociere puternică cu Topicul 3, respectiv Topicul 4. The Picture of Dorian Gray si The Call of the Wild au o asociere puternică cu Topicul 2, ceea ce sugerează că textele acestora conțin subiecte comune. Cartea Frankenstein si Dracula are o asociere cu Topicul 1, insa, spre deosebire de Frankenstein, cartea Dracula are o asociere mai slaba cu acest topic.

Nr.Crt.	title	chapter	topic	gamma
1	The call of the wild	3	1	1
2	The call of the wild	6	1	1
3	The call of the wild	7	1	1
4	The Picture of Dorian Gray	2	1	9.16e-06
5	DRACULA	35	1	8.14e-06
6	DRACULA	39	1	7.43e-06
7	The call of the wild	1	1	1
8	The call of the wild	5	1	1
9	The Picture of Dorian Gray	3	1	1.03e-05
10	The call of the wild	4	1	1

a)



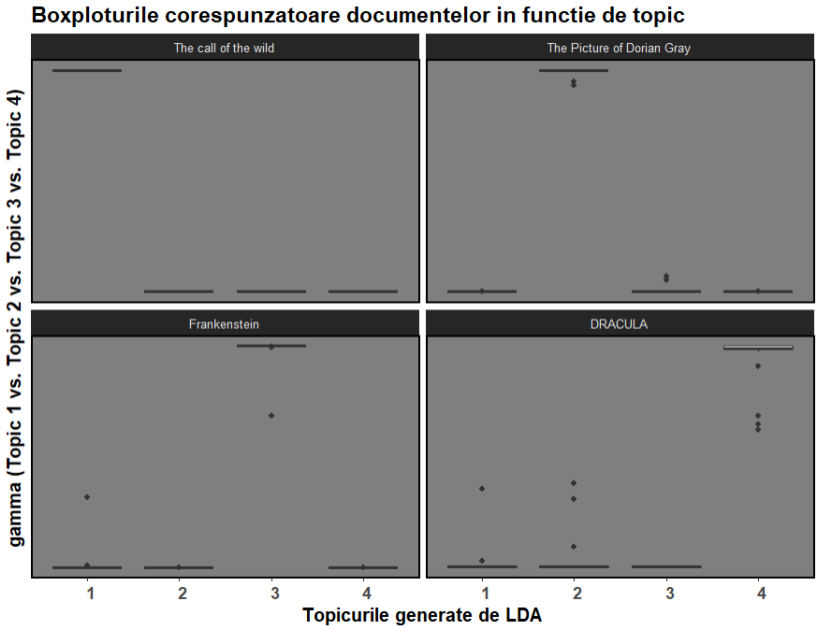
b)

Fig. 3. Asocierea capitolului cartilor analizate cu topicurile create prin modelul LDA a) valoarea indicatorului gamma pentru fiecare carte și b) reprezentarea grafica a distributiei capitolelor din cartile analizate dupa indicatorul gamma in functie de topicul abordat.

In Fig. 3 a) este prezentata asocierea dintre capitolele cărților analizate și subiectul identificat de modelul LDA. Capitolele cărții The Call of the Wild extrase in urma selectiei sunt asociate în mod exclusiv cu Topicul 1 (gamma = 1.00). Aceasta indică faptul că tema principală a cărții este bine reprezentată de Topicul 1, fără influențe semnificative din alte subiecte. Capitolele cărții The Picture of Dorian Gray (12 și 17) au valori gamma foarte mici pentru Topicul 1 fapt ce sugerează că Topicul 1 nu este relevant pentru aceste capitole și că textul acestora este asociat cu alte subiecte (ex.: Topicul 2).

	title	chapter	topic	gamma
	<chr>	<int>	<int>	<dbl>
1	The call of the wild	3	1	1.00
2	The call of the wild	6	1	1.00
3	The call of the wild	7	1	1.00
4	The call of the wild	1	1	1.00
5	The call of the wild	5	1	1.00
6	The call of the wild	4	1	1.00
7	The call of the wild	2	1	1.00
8	The Picture of Dorian Gray	12	1	0.0000213
9	Frankenstein	11	1	0.316
10	The Picture of Dorian Gray	17	1	0.0000256

In Fig. 3 b) este reprezentata asocierea **fiecărui capitol din document (carte)** cu **subiectele principale (Topic 1, 2, 3 și 4)**, in functie de valorile gamma, care indică proporția fiecărui subiect într-un capitol. Observam ca anumite capitole din cartea The Call of the Wild și Frankenstein sunt parțial asociate cu Topicul 1, dar contribuția acestor capitole corespunzatoare cartii Frankenstein este foarte mica (valori gamma foarte mici). Capitolele extrase din cartea The Picture of Dorian Gray are o asociere puternică cu Topicul 2, ceea ce sugerează că acest subiect conține teme comune în acest text, in timp ce capitolele extrase din cartea Frankenstein au o asociere puternică cu Topicul 3, respectiv cele extrase din cartea Dracula au o asociere puternică cu Topicul 4.



In Fig. 4 este prezentata boxploturile corespunzatoare cartilor analizate in functie de topic, indicatorul **gamma** oferind proporția subiectului asociat fiecărui document/carte. Observam ca fiecare carte este asociata exclusiv cu subiectul corespunzator (fiecare carte este asociat cu termeni din topicul corespunzator), pentru celelalte subiecte valorile gamma sunt foarte mici. Prin urmare, exista o asociere clara intre carti si subiecte ceea ce arata ca modelul LDA a reusit sa identifice bine topicul din fiecare carte. Topicurile sunt bine delimitate in functie de tema cartii, ceea ce indica ca textele au fost clasificate in mod correct.

	title	chapter	topic	gamma
	<chr>	<int>	<int>	<dbl>
1	The call of the wild	3	1	1.00
2	The call of the wild	6	1	1.00
3	The call of the wild	7	1	1.00
4	The call of the wild	1	1	1.00
5	The call of the wild	5	1	1.00
6	The call of the wild	4	1	1.00
7	The call of the wild	2	1	1.00
8	The Picture of Dorian Gray	2	2	1.00
9	The Picture of Dorian Gray	3	2	0.949
10	The Picture of Dorian Gray	8	2	1.00

Rezultatele analizei **topic modeling** (realizată cu un model LDA) pe baza capitolelor din cele patru cărți este prezentata mai sus. Fiecare linie oferă informații despre un anumit capitol din carte și **probabilitatea** (gamma) ca acel capitol să aparțină unui anumit **topic**.

Observam faptul ca c**apitolele 1-7** sunt toate asociate cu **Topicul 1** și au o **probabilitate de 1.00**. Aceasta sugerează că textul documentului **The Call of the Wild** este dominant si corespunde unui **singur subiect**, care ar putea fi legat de **natură**, **supraviețuire**, **viața**

sălbatică etc., având în vedere că este o carte despre un câine sălbatic și peripețiile sale. Faptul că toate capitolele (de la 1 la 7) sunt asociate 100% cu acest topic sugerează un subiect dominant pe întreg parcursul cărții, fără schimbări majore în subiect. Fiecare topic este asociat cu o carte dupa cum se poate observa din figura de mai jos.

```
# A tibble: 4 × 2
  consensus                topic
  <chr>                    <int>
1 DRACULA                      4
2 Frankenstein                 3
3 The Picture of Dorian Gray   2
4 The call of the wild         1
```

Mai jos sunt identificati **termenii** care apar fiecarui **capitol**, respectiv se identifica modul in care fiecare termen este asociat cu un **topicul**. Observam că buck și thornton sunt termeni principali în analiza topicului pentru cartea The Call of the Wild, iar aceștia sunt asociați cu topicul 1, ceea ce sugerează ca tema centrală se axează pe personajul principal, Buck, și relațiile sale, în special cu Thornton.

```
# A tibble: 67,667 × 4
   document                term     count .topic
   <chr>                   <chr>    <dbl>  <dbl>
 1 The call of the wild_3  buck        70      1
 2 The call of the wild_6  buck        60      1
 3 The call of the wild_7  buck        56      1
 4 The call of the wild_1  buck        35      1
 5 The call of the wild_5  buck        29      1
 6 The call of the wild_4  buck        34      1
 7 The call of the wild_2  buck        29      1
 8 The call of the wild_6  thornton    50      1
 9 The call of the wild_7  thornton    18      1
10 The call of the wild_5  thornton    13      1
# i 67,657 more rows
```

Mai jos este prezentata asocierea dintre cuvintele fiecarui capitol cu subiectul si cu documentul din care provine. Astfel, tabelul contine urmatoarele variabile: titlul cărții, în acest caz The call of the wild; numărul capitolului din care provine termenul (în capitolul 3, cuvântul buck apare de **70** ori); cuvântul identificat în textul respectiv (cuvintele buck și thornton sunt frecvent întâlnite în acesta carte); frecvența cuvântului în capitolul respectiv (cuvântul buck apare de **70** ori în capitolul 3, iar thornton apare de **50** ori în capitolul 6); subiectul atribuit fiecărui termen (care este **1**, ceea ce sugereaza ca acestia sunt asociați cu **topicul 1** care este tema principală a cărții.); **consensus** arata **titlu** asociat cu fiecare cuvant (termeni buck și thornton) sunt corelați cu acest titlu, ceea ce reflectă faptul că aceștia sunt termeni importanți în cartea The call of the wild.

```
# A tibble: 67,667 × 6
   title                chapter term     count .topic consensus
   <chr>                  <int> <chr>    <dbl>  <dbl> <chr>
 1 The call of the wild       3 buck        70      1 The call of the wild
 2 The call of the wild       6 buck        60      1 The call of the wild
 3 The call of the wild       7 buck        56      1 The call of the wild
 4 The call of the wild       1 buck        35      1 The call of the wild
 5 The call of the wild       5 buck        29      1 The call of the wild
 6 The call of the wild       4 buck        34      1 The call of the wild
 7 The call of the wild       2 buck        29      1 The call of the wild
 8 The call of the wild       6 thornton    50      1 The call of the wild
 9 The call of the wild       7 thornton    18      1 The call of the wild
10 The call of the wild       5 thornton    13      1 The call of the wild
# i 67,657 more rows
```