Lex corpus analysis

Francesco Bailo

2024-06-23 14:37:27.758887

Data

- Total number of documents: 1103
- Total number of DOCX documents: 576
- Total number of PDF documents: 527
- Total number of matched documents (DOCX in PDF): 524 (90.97%)
- Total number of matched documents (PDF in DOCX): 524 (99.43%)

Unmatched documents

DOCXs without a PDFs 2, 6, 17, 30, 34, 48, 51, 56, 73, 102, 126, 142, 143, 144, 187, 198, 218, 219, 233, 234, 236, 237, 246, 248, 251, 274, 276, 294, 298, 311, 312, 316, 327, 360, 362, 374, 376, 381, 399, 406, 420, 423, 445, 467, 501, 509, 537, 546, 551, 569, 573, 578

PDFs without a DOCXs 368, 371, 418

Method

1. Text was read into R with `readtext::readtext()` (Benoit & Obeng, 2023).
2. For PDF documents, text before a match with the regular expression `(DISEGNO|PROPOSTA) DI LEGGE` was removed. This excluded everything that is not part of the body of the bill.
3. For DOCX documents, text before a match with the regular expression `"\\n(Art.|Articolo|ART.|ARTICOLO)([ ]+)?1`, `. (Art.|Articolo|ART.|ARTICOLO)([ ]+)?1`, `^(Art.|Articolo|ART.|ARTICOLO)([ ]+)?1`, `(Art.|Articolo|ART.|ARTICOLO)([ ]+)?1` was removed.
4. The text was split into sentences with `tidytext::unnest_sentences()` (Silge & Robinson, 2016).
5. The text was pre-processed removing sentences that did not contain any letter (`[a-z]`) or that did not contain at least three words (`\\S+`).
6. For each DOCX document, every sentence was then compare to every sentence of the corresponding PDF document using the "SentenceTransformers" framework (Reimers & Gurevych, 2019) and the "sentence-bert-base" sentence-transformers model for Italian (Edoardo Federici, 2022), calculating the cosine similarity of the two sentences.

Examples

Examples: eight levels

doc_id	sentence1	sentence2	score	level
103	le disposizioni della presente legge si applicano alle regioni a statuto speciale e alle province autonome di trento e di bolzano compatibilmente con i rispettivi statuti e le relative norme di attuazione, anche con riferimento alla legge costituzionale 18 ottobre 2001, n.	stampato su carta riciclata ecologica	-0.06	(-0.161,-0.0146]
389	il decreto del ministro della salute 3 luglio 2003, pubblicato nella gazzetta ufficiale pag. 11 n. 211 dell'11 settembre 2003, recante «tabella delle menomazioni alla integrità psicofisica comprese tra 1 e 9 punti di invalidità», è abrogato.	(risarcimento del terzo trasportato).	-0.05	(-0.161,-0.0146]
116	classe ii (materiali strutturanti semplici): fondi di caffè; filtri di tè; foglie; segatura e trucioli di legno vergine non verniciato e non rivestito; paglia; cortecce; sfalci e potature provenienti da città con popolazione inferiore ai 100.000 abitanti o con popolazione superiore ma con superamenti del pm 10 (50 microg/m3) inferiori ai 35 giorni all'anno o previa analisi chimica degli stessi conforme all'allegato ii del dlgs 75/2010; classe iii (materiali complessi): scarti vegetali di cucina post-cottura (con attenzione al contenuto di olio e sale); bucce di agrumi non trattati; piccole quantità di cenere del camino; avanzi di carne, pesce, salumi e formaggi; lettiera di cani e gatti purchè non mista a sostanze non compostabili; foglie di piante resistenti alla degradazione (magnolia, aghi di pino); tessuti naturali e non colorati (ad esempio fibre di canapa, di lino o di cotone grezze); carta non sbiancata. - materiali non compostabili: tutti quelli non compresi nell'elenco precedente.	in caso di mancato ricorso la sanzione è automaticamente comminata.	0.06	(-0.0146,0.13]

(continued)

doc_id	sentence1	sentence2	score	level
378	per le operazioni di acquisto prodotto effettuate senza l'applicazione della ritenuta ai sensi del comma 4, il soggetto acquirente emette un documento d'acquisto dal quale risulti la data di cessione, nome e cognome, codice fiscale del cedente, codice ricevuta del versamento dell'imposta sostitutiva al comma 1, natura e quantità del prodotto ceduto, nonché l'ammontare del corrispettivo pattuito.	capo ii raccolta dei tartufi art.	0.10	(-0.0146,0.13]
119	non è ammessa la richiesta di verifica di assoggettabilità o di valutazione impatto ambientale in sanatoria.	al comma 1 dell'articolo 32-bis del decreto legislativo 3 aprile 2006, n. 152, e successive modificazioni, le parole: « negativi e », ovunque ricorrono, sono soppresse.	0.19	(0.13,0.275]
106	qualora dal fatto derivi pericolo per l'igiene e la sanità pubblica, il sindaco vieta immediatamente la prosecuzione dell'attività.	è consentito l'utilizzo dei seguenti prodotti tradizionalmente usati a fini conservativi: sale, pepe, zucchero, olio, aceto e simili.	0.23	(0.13,0.275]
220	salvo che il fatto costituiscareato, chiunque viola le disposizioni di cui al comma 5 dell'articolo 1 ter è soggetto alla sanzione amministrativa pecuniaria da euro 20.000a 120.000.	il presente decreto prevede la disciplina sanzionatoria per le violazioni delle norme in materia di designazione e definizione delle categorie degli oli di oliva e degli oli di sansa d'oliva, delle disposizioni del regolamento di esecuzione (ue) n. 29/2012 della commissione, del 13 gennaio 2012, relativo alle norme di commercializzazione dell'olio d'oliva, e del regolamento (cee) n. 2568/91 della commissione, dell'11 luglio 1991, relativo alle caratteristiche degli oli d'oliva e degli oli di sansa d'oliva nonché ai metodi ad essi attinenti ».	0.41	(0.275,0.42]
243	l'agenzia deve essere dotata di un bilancio che le consente di svolgere le sue funzioni, concordemente all' art.4 comma 6 lettera d) del regolamento eu 996/2010.	l'agenzia pubblica annualmente e trasmette entro il 30 settembre alla presidenza del consiglio dei ministri una rela zione sulle attività svolte nell'anno precedente.	0.40	(0.275,0.42]

(continued)

doc_id	sentence1	sentence2	score	level
318	l'azienda unità sanitaria locale comunica i punti di prelievo fissati per il controllo, le frequenze dei campionamenti e gli eventuali aggiornamenti alla competente regione o provincia autonoma ed al ministero della salute secondo modalità proposte dal ministero della salute e sullequali la conferenza stato-regioni esprime intesa, e trasmette gli eventuali aggiornamenti entro trenta giorni dalle variazioni apportate.	è vietato l'approvvigionamento idropotabile da falde contaminate in cui, per una o più sostanze, siano superate le concentrazioni soglia di contaminazione individuate nell'allegato 5 al titolo v della parte quarta del decreto legislativo 3 aprile 2006, n. 152, e successive modificazioni, e comunque da falde incluse nei perimetri di siti inquinati o poten zialmente inquinati e nelle aree immediatamente a valle, in senso idrogeologico, di queste, definite dall'articolo 4-bis del presente decreto.	0.52	(0.42,0.565]
560	a decorrere dall'anno 2018 il piano straordinario di intervento per lo sviluppo del sistema territoriale dei servizi socio-educativi di cui all'articolo 1, comma 1259, della legge 27 dicembre 2006, n. 296, al quale concorrono gli asili nido, i servizi integrativi, diversificati per modalità strutturali, di accesso, di frequenza e di funzionamento, e i servizi innovativi nei luoghi di lavoro, presso le famiglie, ivi compresi i nidi domiciliari, gli asili nido familiari organizzati dalle famiglie, in forma singola o associata, presso il proprio domicilio, gli asili nido di caseggiato organizzati dalle famiglie, è rifinanziato per 100 milioni di euro annui.	la detrazione di cui al presente comma spetta nella misura del 50 per cento delle spese sostenute ed è calcolata su un ammontare complessivo non superiore a 10.000 euro annui.	0.43	(0.42,0.565]

(continued)

doc_id	sentence1	sentence2	score	level
9	con uno o piu‘ decreti del ministro dell´ambiente e della tutela del territorio e del mare, di concerto con il ministro della salute, sono individuate le opzioni che garantiscano l’immissione sul mercato di materiali e prodotti concepiti in modo da ridurre il più possibile, in relazione alla loro realizzazione, nonché al riuso o allo smaltimento di essi , la quantità o la nocività dei rifiuti e i rischi di inquinamento nell’ottica della eco-progettazione di cui all’art.179-ter.	devono inoltre essere redatti programmi di prevenzione dei rifiuti basati sulle migliori pratiche, tenendo conto dell’analisi del ciclo di vita di cui al citato articolo 179-ter nelle decisioni sulle misure da adottare »; e) al comma 5: 1) all’alinea, dopo le parole: « le pubbliche amministrazioni » sono inserite le seguenti: « , supportate dai centri di ricerca appositamente istituiti, »; 2) la lettera b) è abrogata; 3) la lettera e) è sostituita dalla seguente: « e) l’utilizzo di metodologie di analisi del ciclo di vita di cui all’articolo 179-ter quale strumento decisionale a supporto delle attività di trattamento dei rifiuti »; f) il comma 6 è sostituito dal seguente: « 6.	0.68	(0.565,0.71]
398	il giudice tributario onorario in pubblica udienza o il collegio in camera di consiglio, sentite le parti e delibato il merito, provvede con ordinanza motivata non impugnabile.	al magistrato tributario onorario eletto spetta il trattamento economico percepito dal magistrato tributario eletto con la medesima valutazione di professionalità o, in mancanza, con quella immediatamente superiore.	0.60	(0.565,0.71]
389	12 (sinistri verificatesi nel territorio della repubblica) al comma secondo dell’articolo 283 del decreto legislativo 7 settembre 2005, n. 209, sono apportate le seguenti modifiche: 1. al primo periodo, le parole da “il risarcimento è dovuto solo per i danni alla persona” a “nei casi di cui al comma 1 lettere” sono soppresse; 2. dopo il secondo periodo è aggiunto il seguente terzo periodo: “nel caso di cui al comma 1 lettera a) il risarcimento per i danni alle cose è dovuto in misura non superiore a euro 10.000».	al comma 1 dell’articolo 146 del codice di cui al decreto legislativo 7 settembre 2005, n. 209, dopo le parole: « ai contraenti ed ai danneggiati » sono inserite le seguenti: « nonché ai titolari delle imprese che provvedono alle riparazioni dei veicoli coinvolti nel sinistro muniti di delega, o di cessione di credito, ».	0.72	(0.71,0.855]

(continued)

doc_id	sentence1	sentence2	score	level
9	all'art.178 sono apportate le seguenti modifiche: a) al comma 1, dopo le parole: "fattibilità tecnica ed economica"sono inserite le seguenti: "e responsabilità estesa del produttore del prodotto" b) dopo il comma 1, sono aggiunti, in fine, i seguenti: "2.	all'articolo 198 del decreto legislativo n. 152 del 2006, sono apportate le seguenti modificazioni: a) al comma 1: 1) le parole: « ed assimilati » sono soppresse; 2) il secondo periodo è sostituito dal seguente: « fino all'inizio delle attività del soggetto aggiudicatario della gara ad evidenza pubblica indetta dall'autorità d'ambito ai sensi dell'articolo 202, i comuni continuano la gestione dei rifiuti pericolosi e non pericolosi urbani e dei rifiuti assimilati avviati allo smaltimento in regime di privativa nelle forme previste dal regolamento di cui al decreto del presidente della repubblica 7 settembre 2010, n. 168 »; b) al comma 2: 1) all'alinea, le parole: « princìpi di trasparenza, efficienza, efficacia ed economicità » sono sostituite dalle seguenti: « princìpi di trasparenza, efficienza, efficacia e sostenibilità ambientale »; 2) alla lettera c), le parole: « ed assimilati » sono soppresse; 3) alla lettera e), dopo le parole: « le forme di » è inserita la seguente: « prevenzione, »; 4) alla lettera g), dopo le parole: « secondo i criteri » sono inserite le seguenti: « e le limitazioni »; c) il comma 3 è sostituito dal seguente: « 3.	0.77	(0.71,0.855]
562	quando gli atti o i documenti sono stati assoggettati al vincolo di segreto funzionale da parte delle competenti commissioni parlamentari di inchiesta, tale segreto non può essere opposto alla commissione.	quando gli atti o i documenti sono stati assoggettati al vincolo del segreto funzionale da parte delle competenti commissioni parlamentari di inchiesta, tale segreto non può essere opposto alla commissione.	1.00	(0.855,1]

(continued)

doc_id	sentence1	sentence2	score	level
335	il ministero dello sviluppo economico, nel rispetto ed in coordinamento con la normativa comunitaria e nazionale vigente in materia, promuove iniziative di sensibilizzazione dei media, volte ad una maggiore tutela dei minori nei riguardi della pubblicità ad essi rivolta, che informi sulla pericolosità degli alimenti e/o bevande contenenti un elevato apporto totale di lipidi per porzione, grassi trans, oli vegetali, zuccheri semplici aggiunti, alto contenuto di sodio, nitriti e/o nitrati utilizzati come additivi, aggiunta di zuccheri semplici e dolcificanti, elevato contenuto di teina, caffeina, taurina e similari.	il ministero dello sviluppo economico, nel rispetto e in coordinamento con la normativa nazionale e dell'unione europea vigente, promuove iniziative di sensibilizzazione dei mezzi di informazione volte a una maggiore tutela dei minori nei riguardi della pubblicità a essi rivolta, prevedendo la diffusione di notizie sui rischi legati al consumo degli alimenti e delle bevande contenenti un elevato apporto totale per porzione di lipidi, di grassi transgenici, di oli vegetali, di zuccheri semplici aggiunti, di sodio, di nitriti o di nitrati utilizzati come additivi, di zuccheri semplici e di dolcificanti, di teina, di caffeina, di taurina e di sostanze simili.	0.94	(0.855,1]

Examples: >.95 (high similarity)

doc_id	sentence1	sentence2	score	level
328	ai datori di lavoro privati che nei casi di accoglimento delle richieste di cui al comma 1 attivano nuovi contratti di lavoro a tempo parziale, senza alcun licenziamento nell'arco dei sei mesi precedenti e dei 18 mesi successivi la nuova assunzione, è riconosciuto, ferma restando l'aliquota di computo delle prestazioni pensionistiche, l'esonero dal versamento dei complessivi contributi previdenziali a carico dei datori di lavoro, per un periodo massimo di 24 mesi, con esclusione dei premi e contributi dovuti all'inail.	ai datori di lavoro privati che nei casi di accoglimento delle richieste di cui al comma 1 attivano nuovi contratti di lavoro a tempo parziale, senza alcun licenziamento nell'arco dei sei mesi precedenti e dei diciotto mesi successivi la nuova assunzione, è riconosciuto, ferma restando l'aliquota di computo delle prestazioni pensionistiche, l'esonero dal versamento dei complessivi contributi previdenziali e assicura tivi a carico dei datori di lavoro, per un periodo massimo di ventiquattro mesi.	0.99	(0.855,1]

(continued)

doc_id	sentence1	sentence2	score	level
184	la commissione, al termine dei lavori e comunque ogniqualvolta lo ritenga necessario, riferisce alle camere sui risultati della propria attività e formula osservazioni e proposte sugli effetti, sui limiti e sull'eventuale necessità di adeguamento della normativa vigente.	la commissione, al termine dei lavori e comunque ogniqualvolta che lo ritiene necessario, riferisce alle camere sui risultati della propria attività e formula osservazioni e proposte sugli effetti, sui limiti e sull'eventuale necessità di adeguamento della normativa vigente.	1.00	(0.855,1]
305	all'onere derivante dall'attuazione della presente legge, paria a 40 milioni di euro annui per l'anno 2019 e a 80 milioni a decorrere dall'anno 2020, si provvede mediante corrispondente riduzione dello stanziamento del fondo speciale di parte corrente iscritto, ai fini del bilancio triennale 2019-2021, nell'ambito del programma «fondi di riserva e speciali» della missione «fondi da ripartire» dello stato di previsione del ministero dell'economia e delle finanze per l'anno 2019, allo scopo parzialmente utilizzando l'accantonamento relativo al medesimo ministero.	all'onere derivante dall'attuazione della presente legge, pari a 40 milioni di euro per l'anno 2019 e a 80 milioni di euro annui a decorrere dall'anno 2020, si provvede me diante corrispondente riduzione dello stan ziamento del fondo speciale di parte cor rente iscritto, ai fini del bilancio triennale 2019-2021, nell'ambito del programma « fondi di riserva e speciali » della missione « fondi da ripartire » dello stato di previ sione del ministero dell'economia e delle fi nanze per l'anno 2019, allo scopo parzial mente utilizzando l'accantonamento relativo al medesimo ministero.	0.98	(0.855,1]
137	ogni osservatorio regionale (o interregionale), è composto, in forma paritetica, da docenti universitari, dirigenti sanitari delle strutture presso le quali si svolge la formazione, dai coordinatori dei poli formativi dei corsi di formazione specifica di medicina generale, nonché da rappresentanti dei medici in formazione, specialistica e specifica di medicina generale.	ogni osservatorio regionale o interregionale è composto, in numero uguale, da docenti universitari, da dirigenti sanitari delle strutture presso le quali si svolgono le attività di formazione di cui al comma 1, dai coordinatori dei poli formativi dei corsi di formazione specifica di medicina generale, nonché da rappresentanti dei profili medici e non medici in formazione specialistica e specifica di medicina generale.	0.98	(0.855,1]

(continued)

doc_id	sentence1	sentence2	score	level
254	entro tre mesi dalla pubblicazione del documento di cui al presente comma, il proponente dichiara pubblicamente se intende: a) rinunciare al progetto o presentarne uno alternativo; b) proporre modifiche al progetto, indicando quelle che intende realizzare; c) sostenere il medesimo progetto sul quale si è svolto il dibattito pubblico, argomentando le ragioni di tale scelta.	entro tre mesi dalla pubblicazione del documento di cui al presente comma, il proponente dichiara pubblicamente se intende: a) rinunciare al progetto o presentarne uno alternativo; b) proporre modifiche al progetto, indicando quelle che intende realizzare; c) sostenere il medesimo progetto sul quale si è svolto il dibattito pubblico, argomentando le ragioni di tale scelta.	1.00	(0.855,1]
517	5 (poteri e limiti della commissione) 1.	(poteri e limiti della commissione) 1.	0.96	(0.855,1]
506	in caso di parità di voti è proclamato eletto o entra in ballottaggio il più anziano di età.	in caso di parità di voti è proclamato eletto o entra in ballottaggio il più anziano di età.	1.00	(0.855,1]
473	il ministro dell'economia e delle finanze può indicare, nel suddetto decreto, il presidente del consiglio di amministrazione.	il ministro dell'economia e delle finanze può indicare, nel suddetto decreto, il presidente del consiglio di amministrazione.	1.00	(0.855,1]

Summary of results

- Number of DOCX documents: 516
- Average proportion of high sentence-sentence similarity: 54.9%
- Median proportion of high sentence-sentence similarity: 58.33%
- Proportion of DOCX with less than 20% of high-similarity sentences: 83 (16.09)%
- Proportion of DOCX with more than 50% of high-similarity sentences: 297 (57.56)%
- Proportion of DOCX with more than 80% of high-similarity sentences: 107 (20.74)%
- Proportion of DOCX with more than 90% of high-similarity sentences: 57 (11.05)%
- Mean sentence similarity (of most similar pairs):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5248  0.8636  0.9210  0.8951  0.9547  1.0000

* Mean: 0.9

* Standard deviation: 0.09

* Least similar 5%:

##   5%
## 0.72
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5248  0.5671  0.6212  0.6247  0.6965  0.7150
```

- Token/word cosine similarity (cosine similarity with tf-idf for robustness):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01918 0.65215 0.80065 0.75417 0.90617 0.99605
```

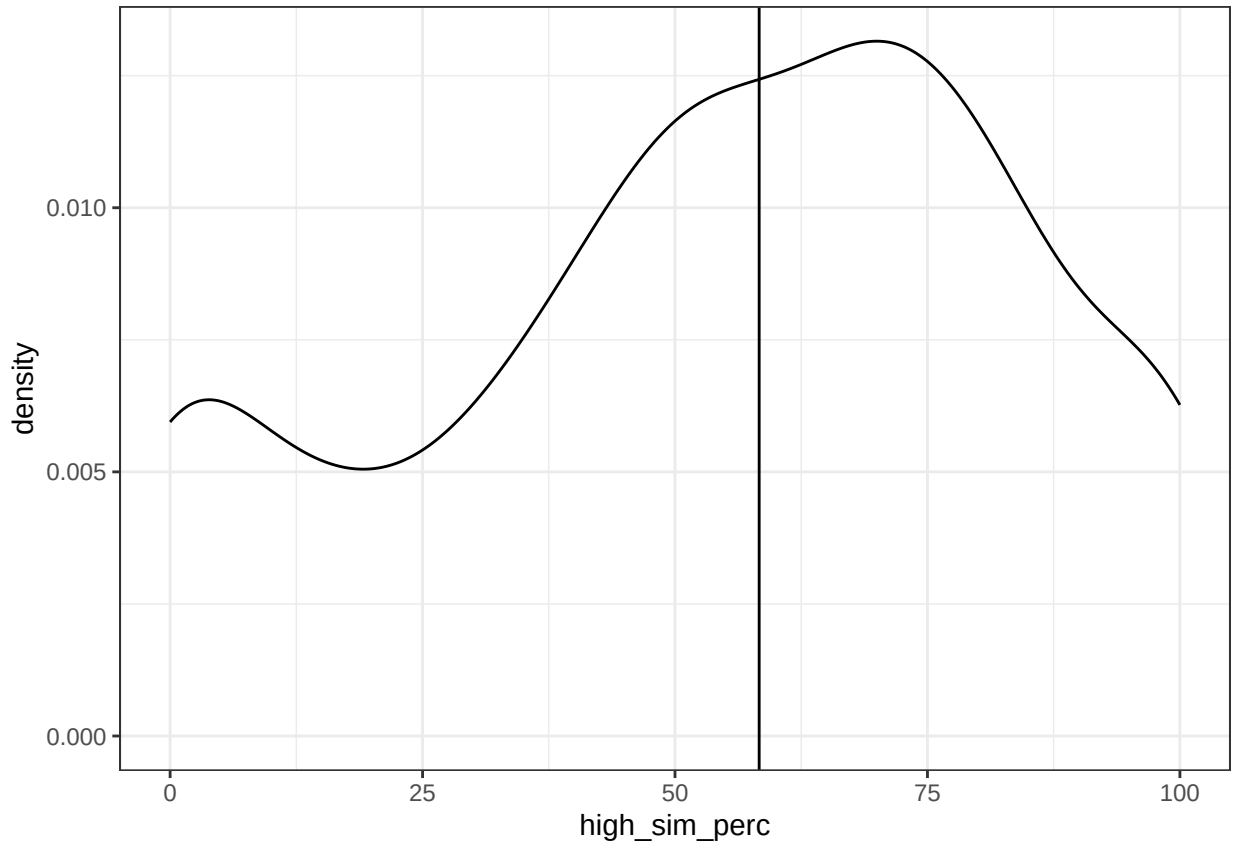


Figure 1: High similarity sentences by DOCX (with median)

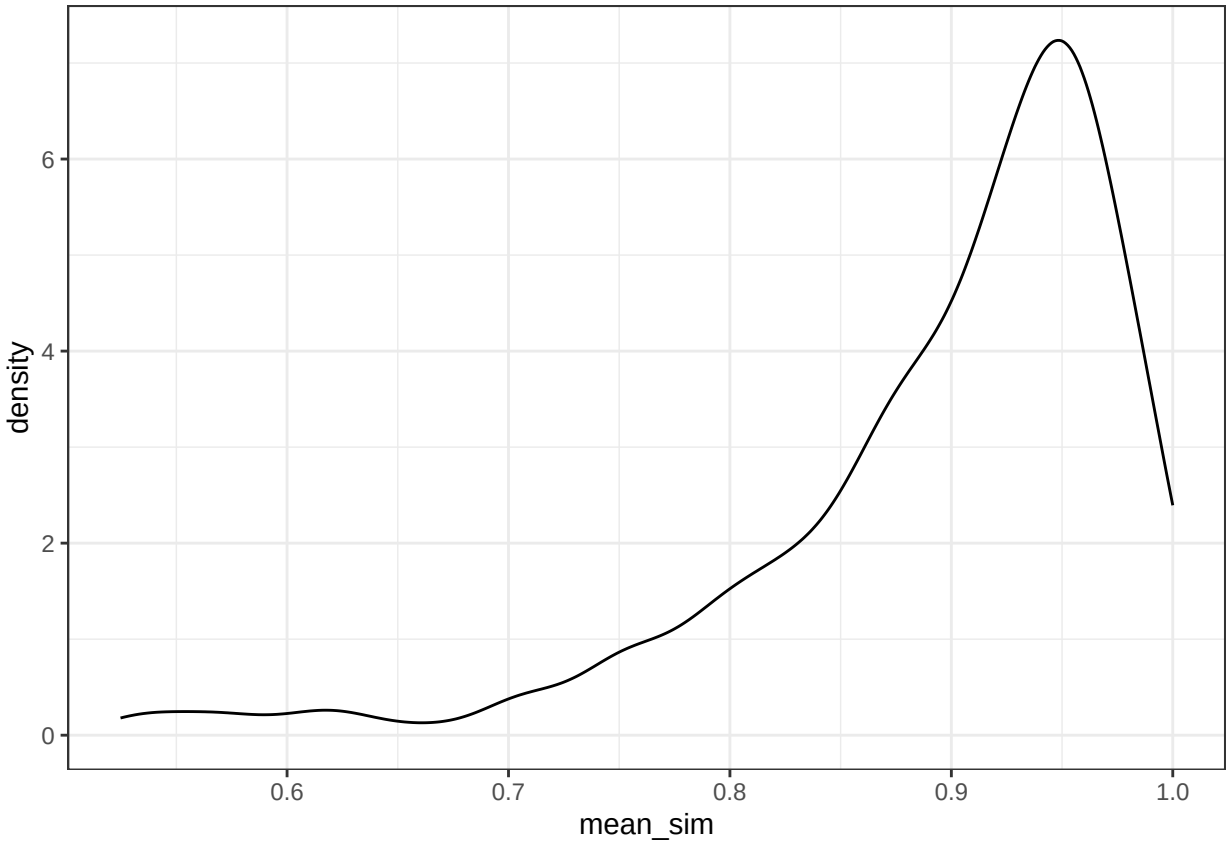


Figure 2: Mean similarity score by DOCX

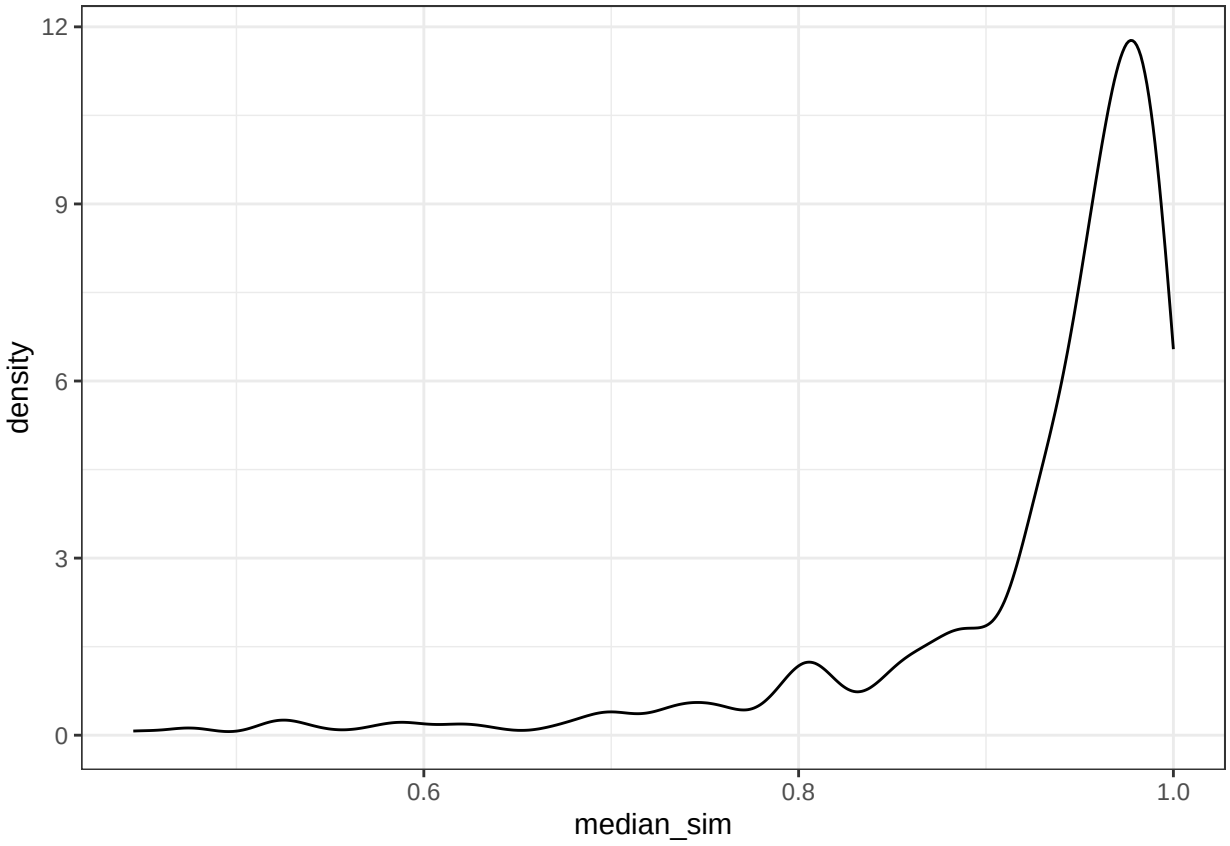


Figure 3: Median similarity score by DOCX

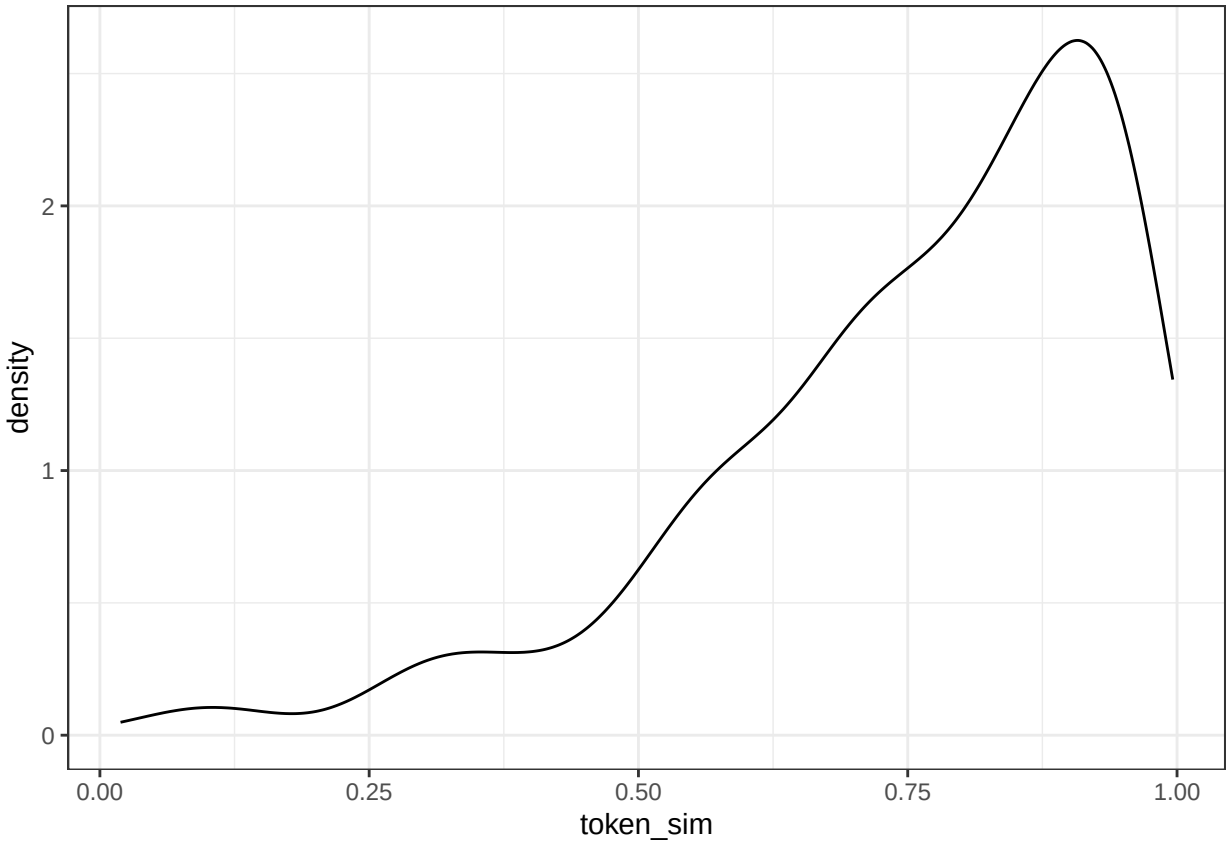
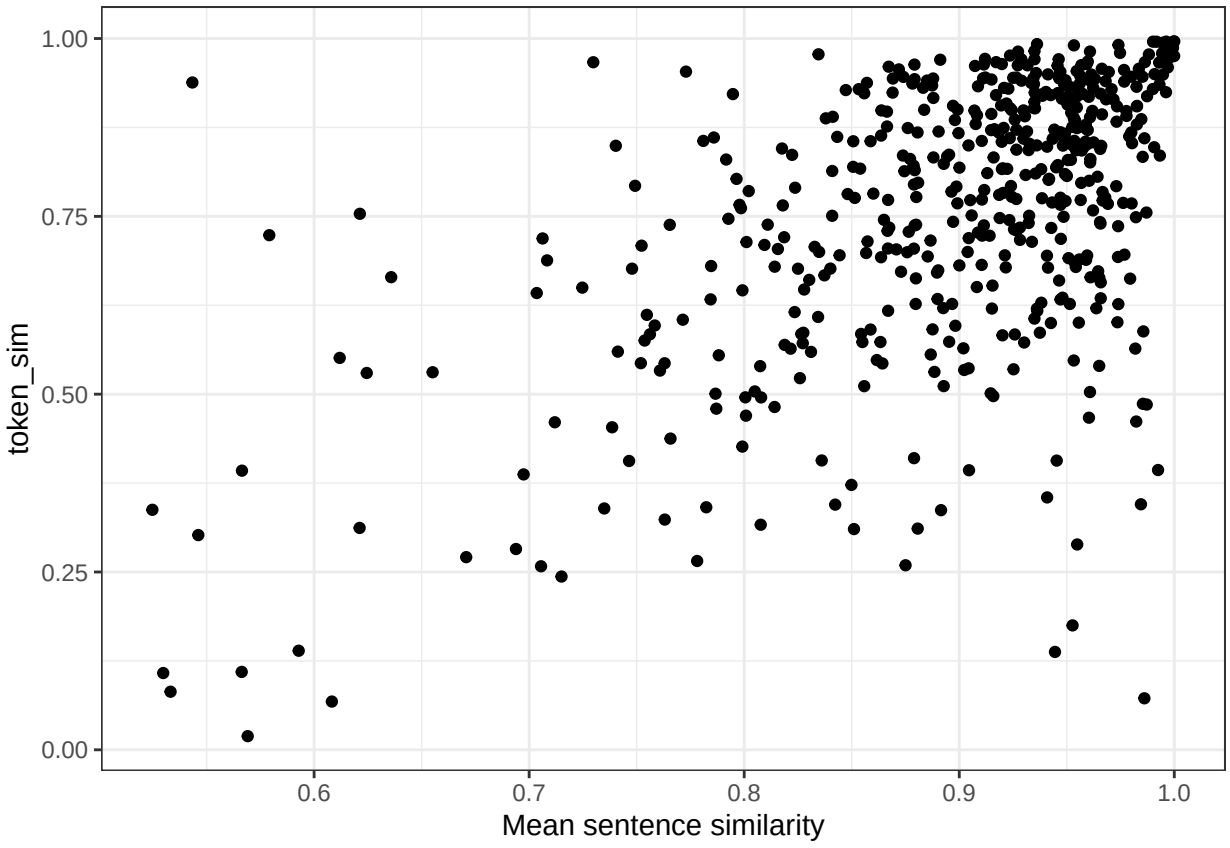


Figure 4: Token cosine similarity



quantile	0+ responses	1+ responses	5+ responses	mean responses
bottom 5%	0.3462	0.3077	0.1538	15.1923
bottom 5-10%	0.1154	0.1154	0.1154	2.5000
top 90%	0.1250	0.1164	0.0948	2.2263

0+ responses	1+ responses	5+ responses	mean responses
0.1357	0.126	0.0988	2.8934

Regression

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Jun 23, 2024 - 14:37:39

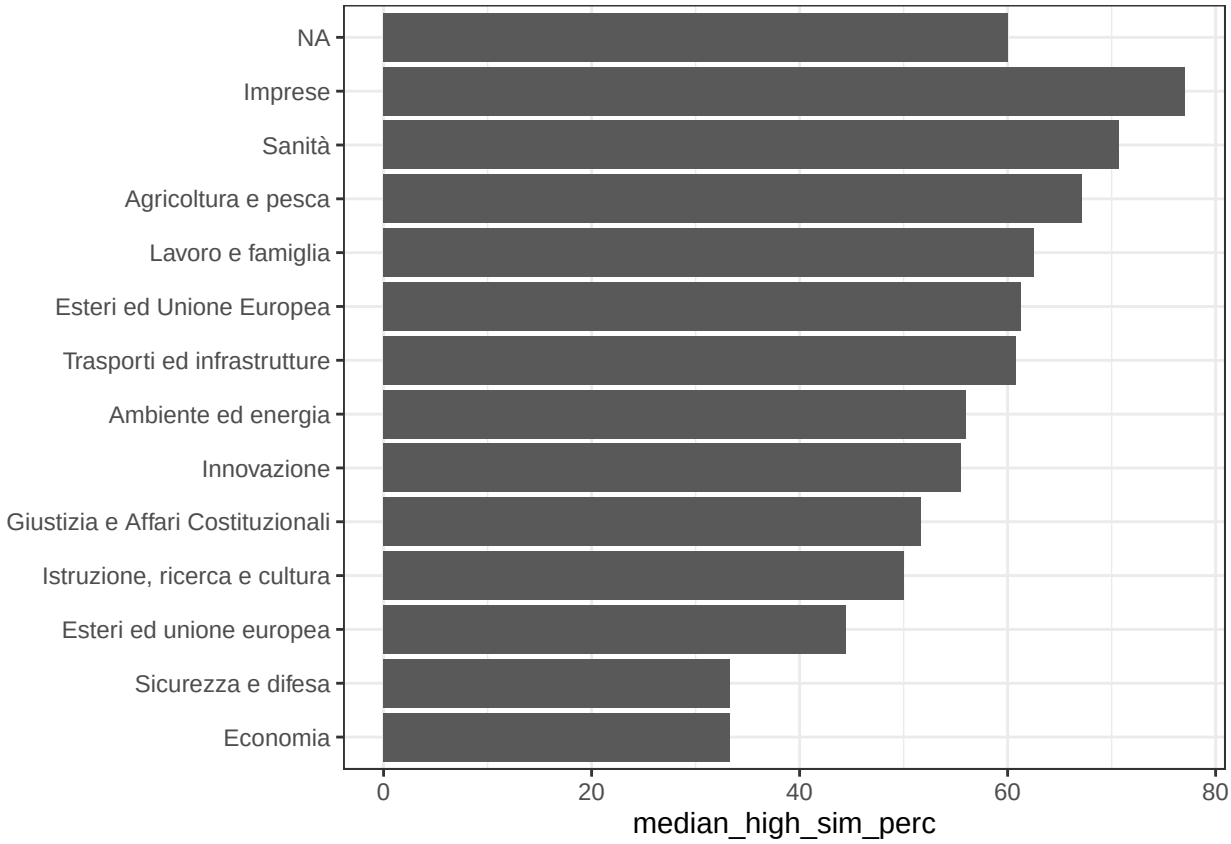


Figure 5: High similarity sentences by Tematica (median)

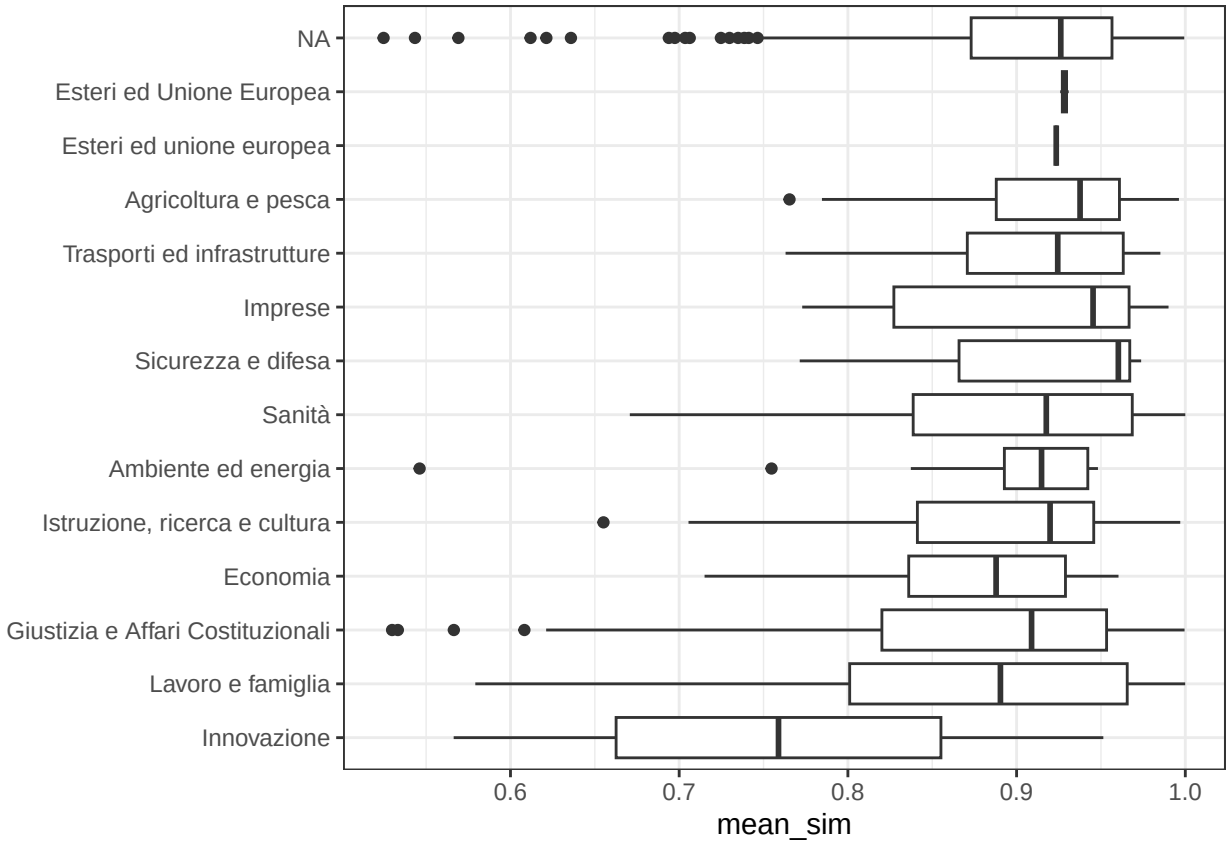


Figure 6: Mean similarity by Tematica

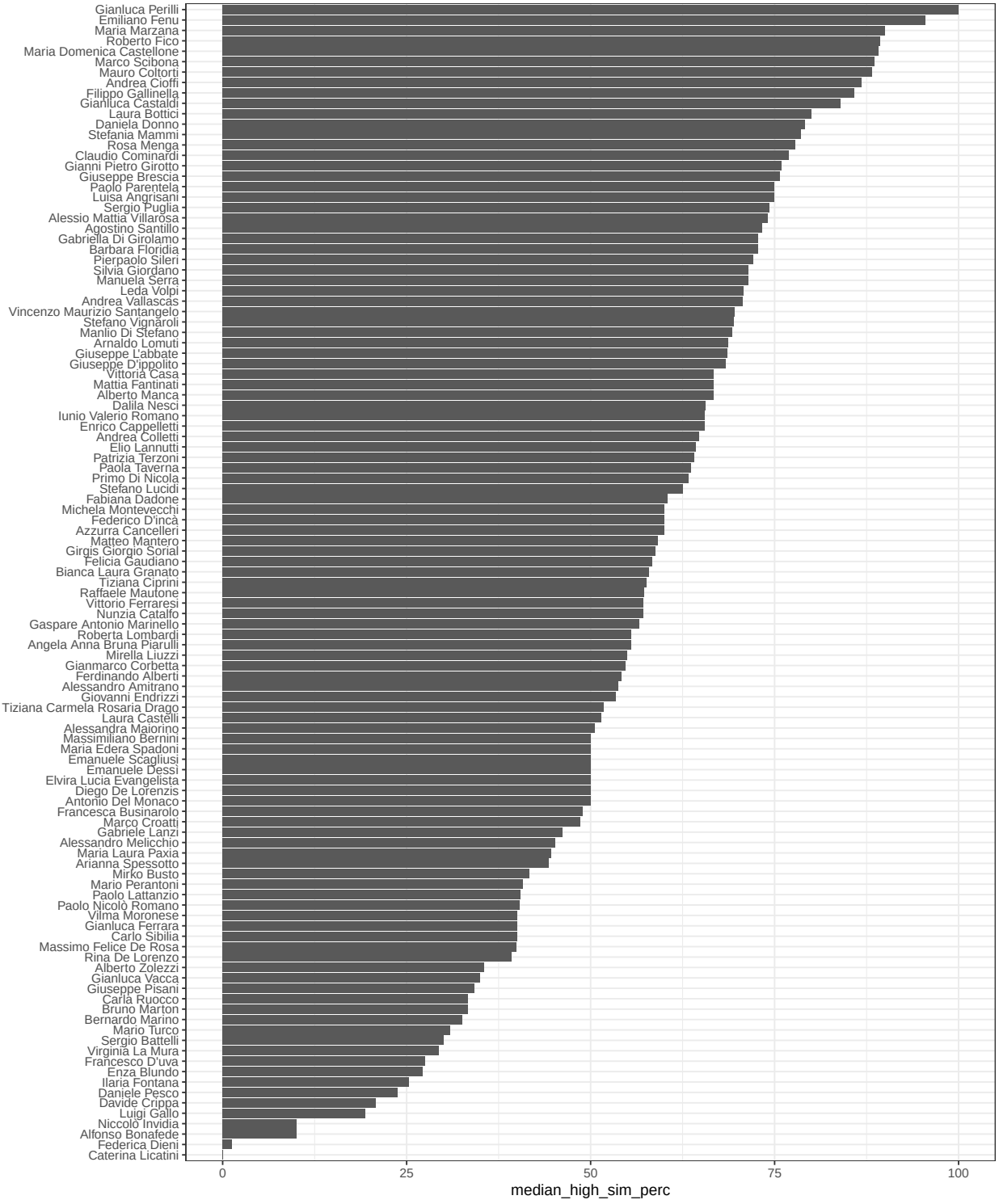


Figure 7: High similarity sentences by firmatario (median, only if Firmatario has more than one proposal)

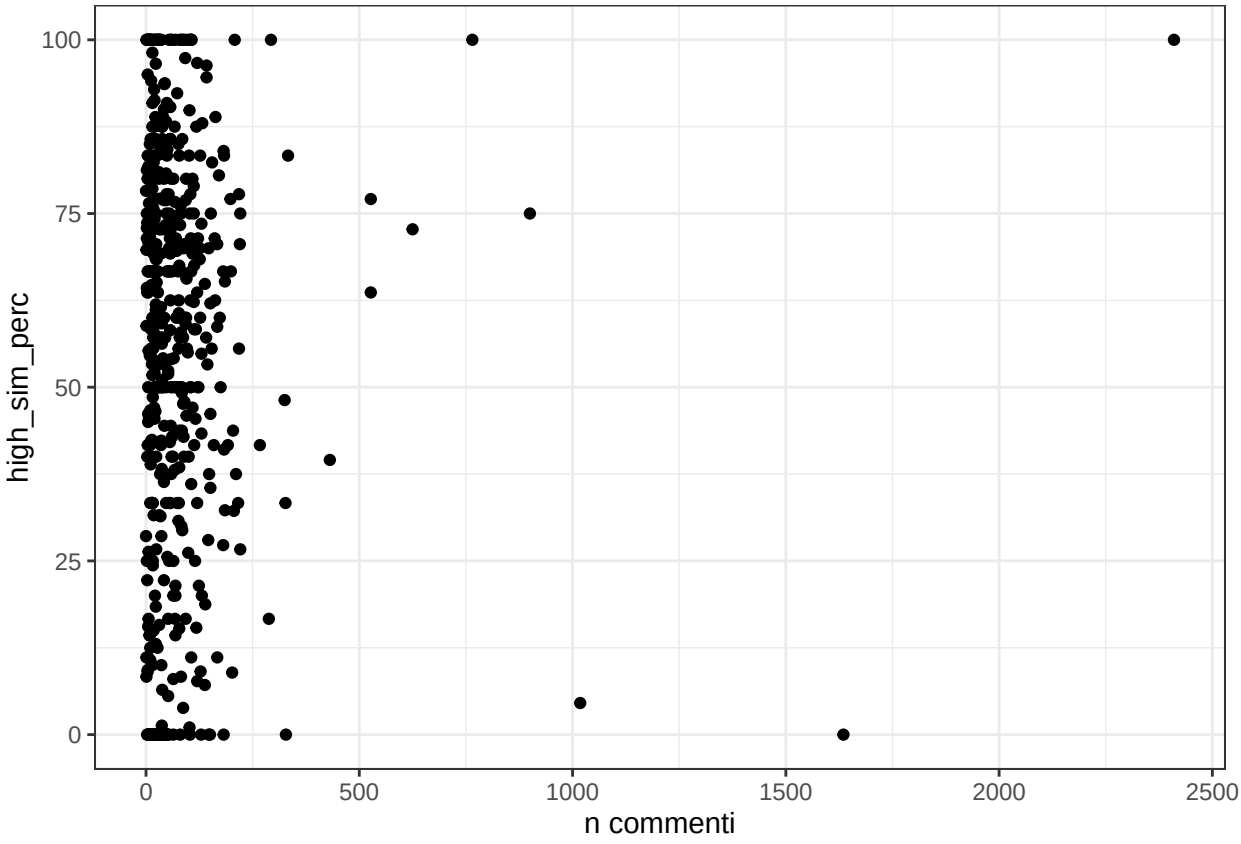


Figure 8: High similarity (%) by n commenti

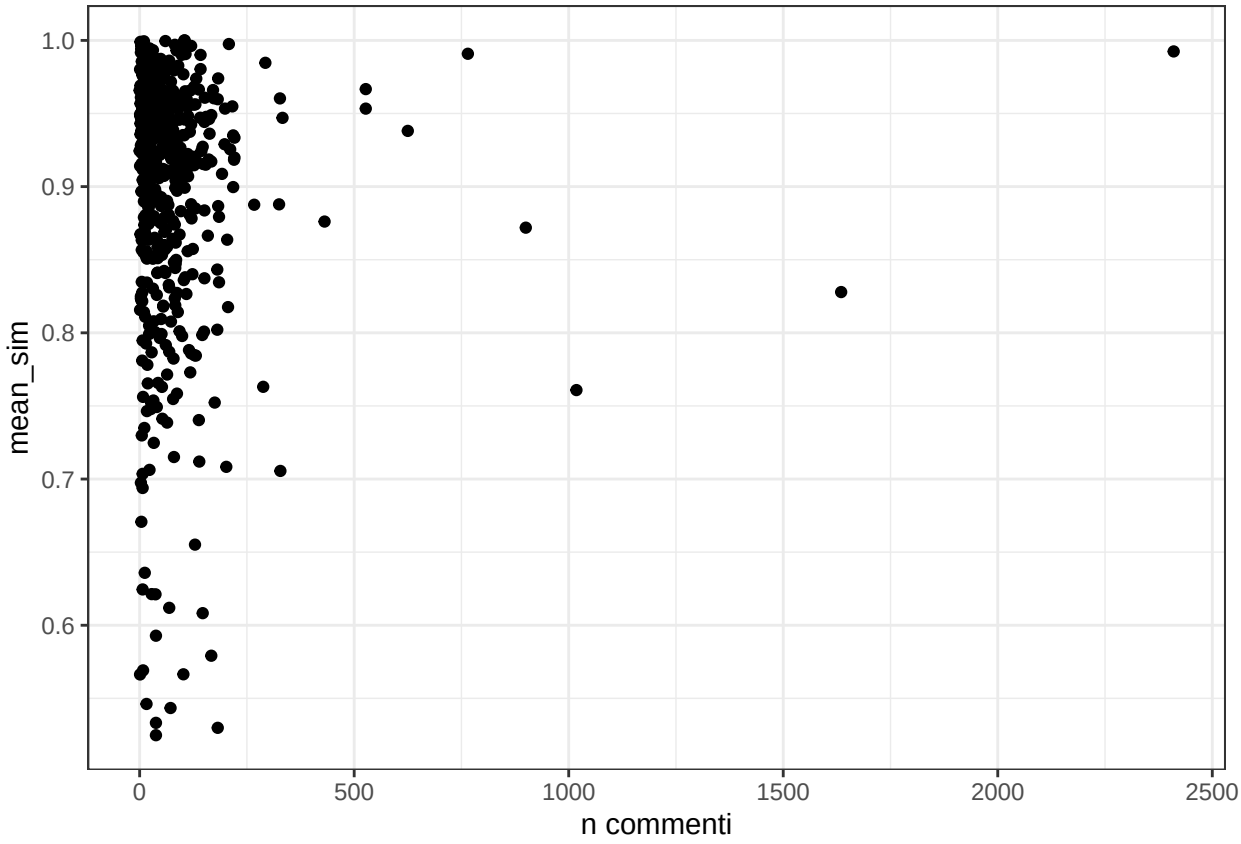


Figure 9: Mean similarity by n commenti

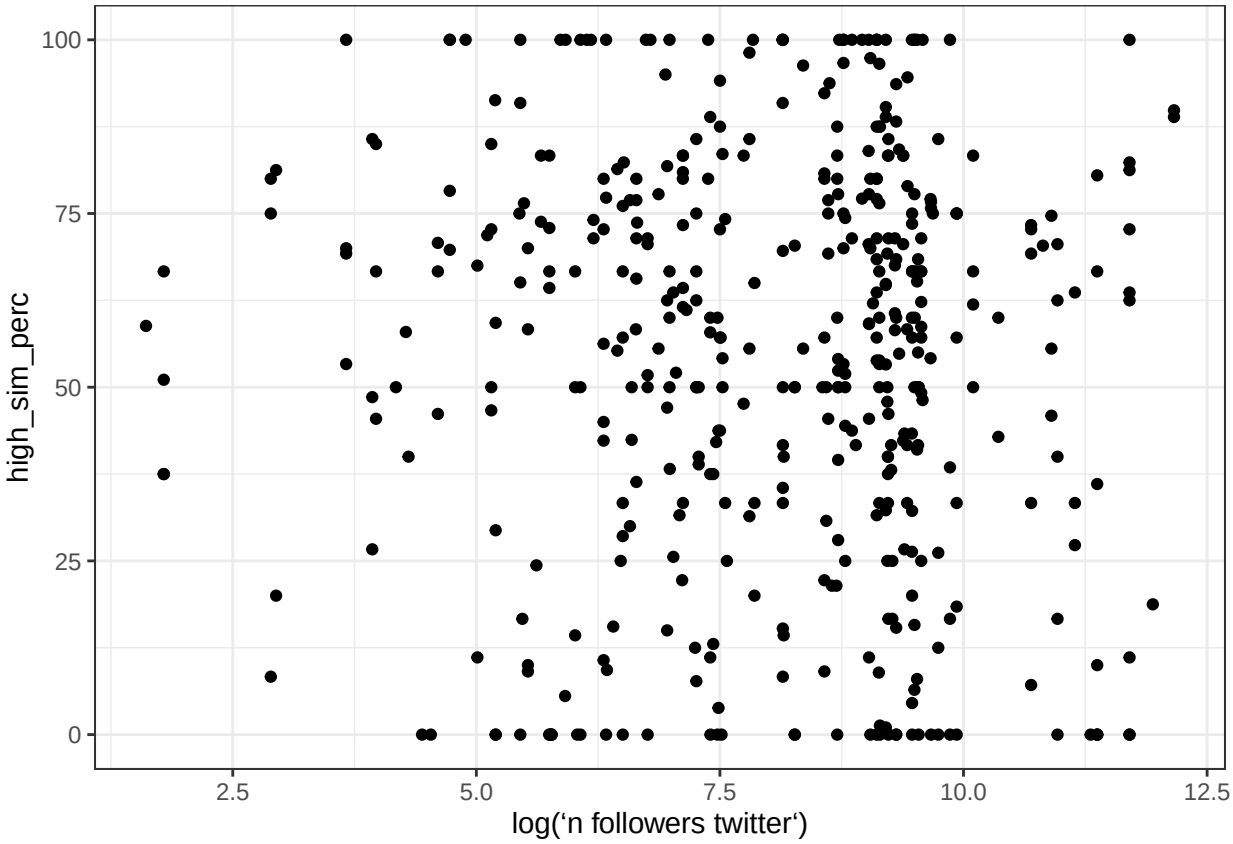


Figure 10: High similarity (%) by n followers twitter (log)

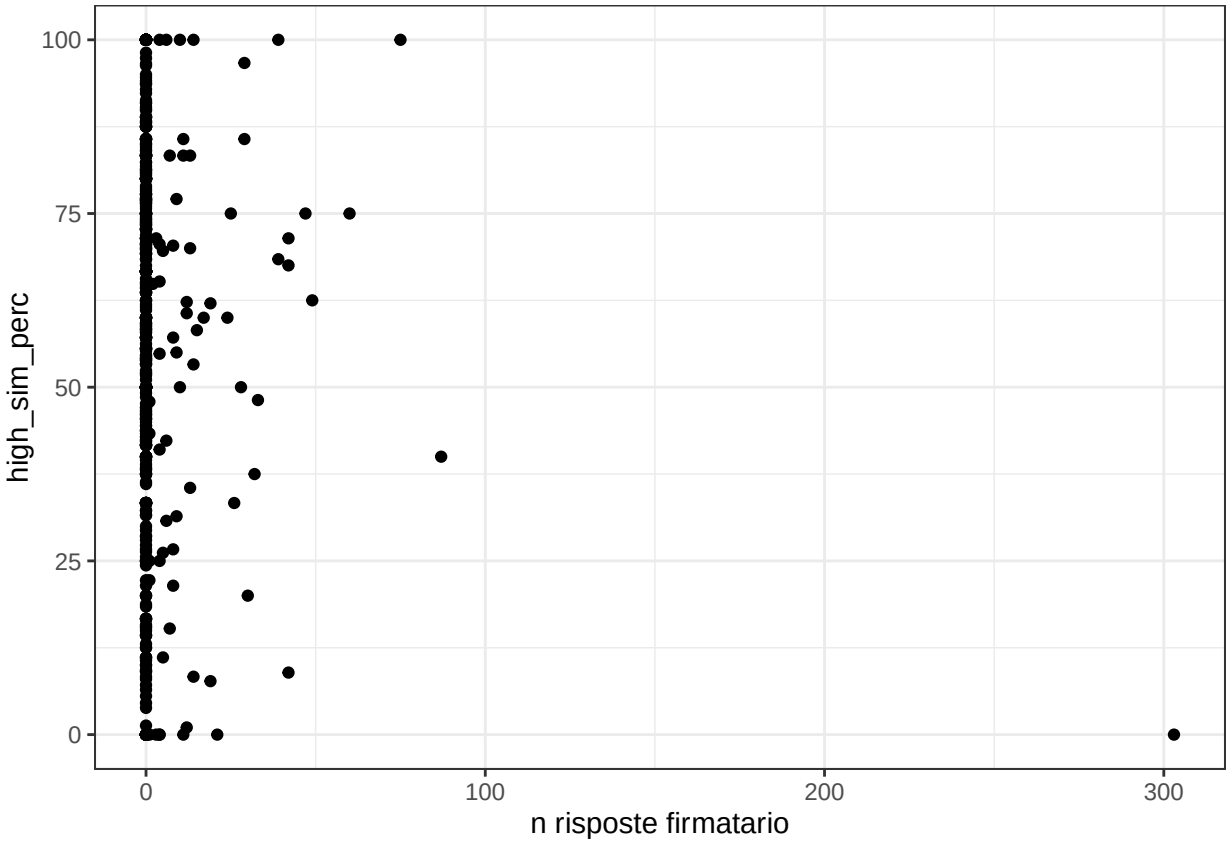


Figure 11: High similarity (%) by n risposte firmatario

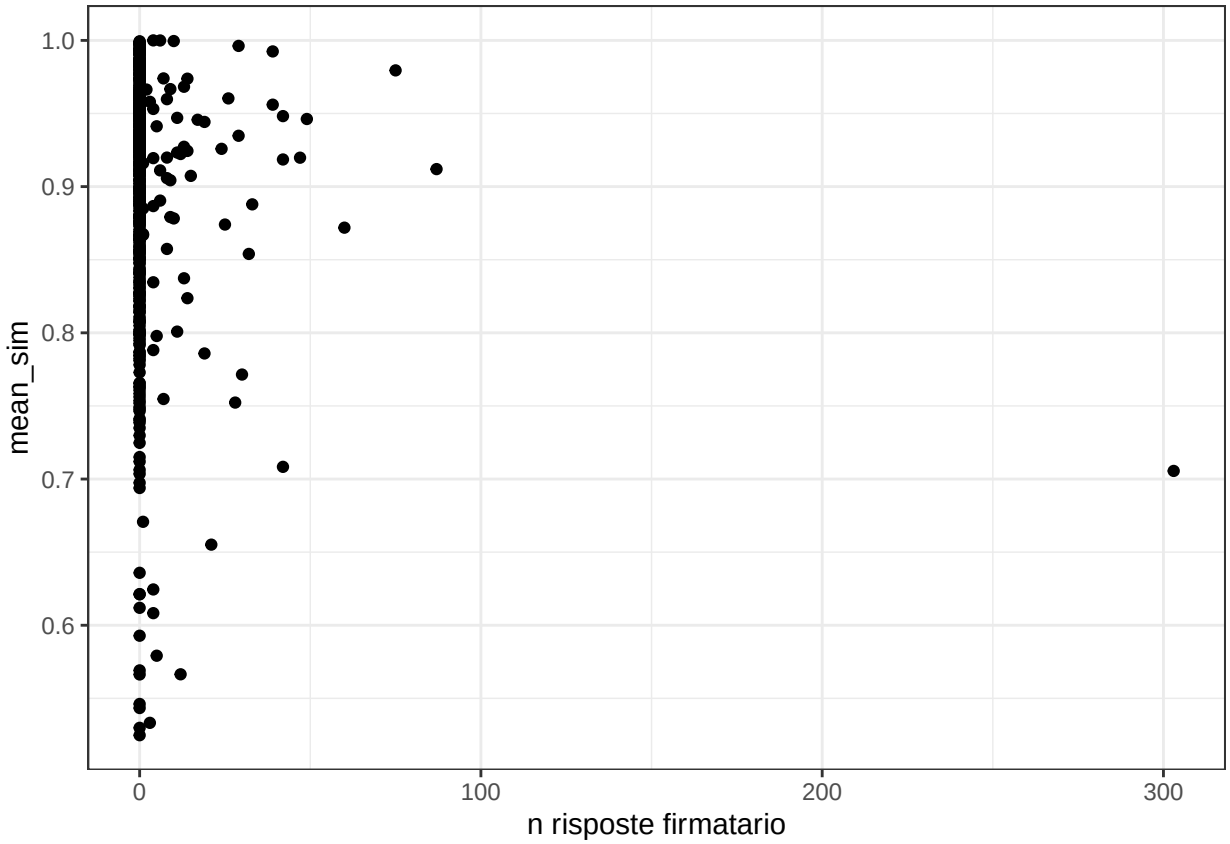


Figure 12: Mean similarity by n risposte firmatario

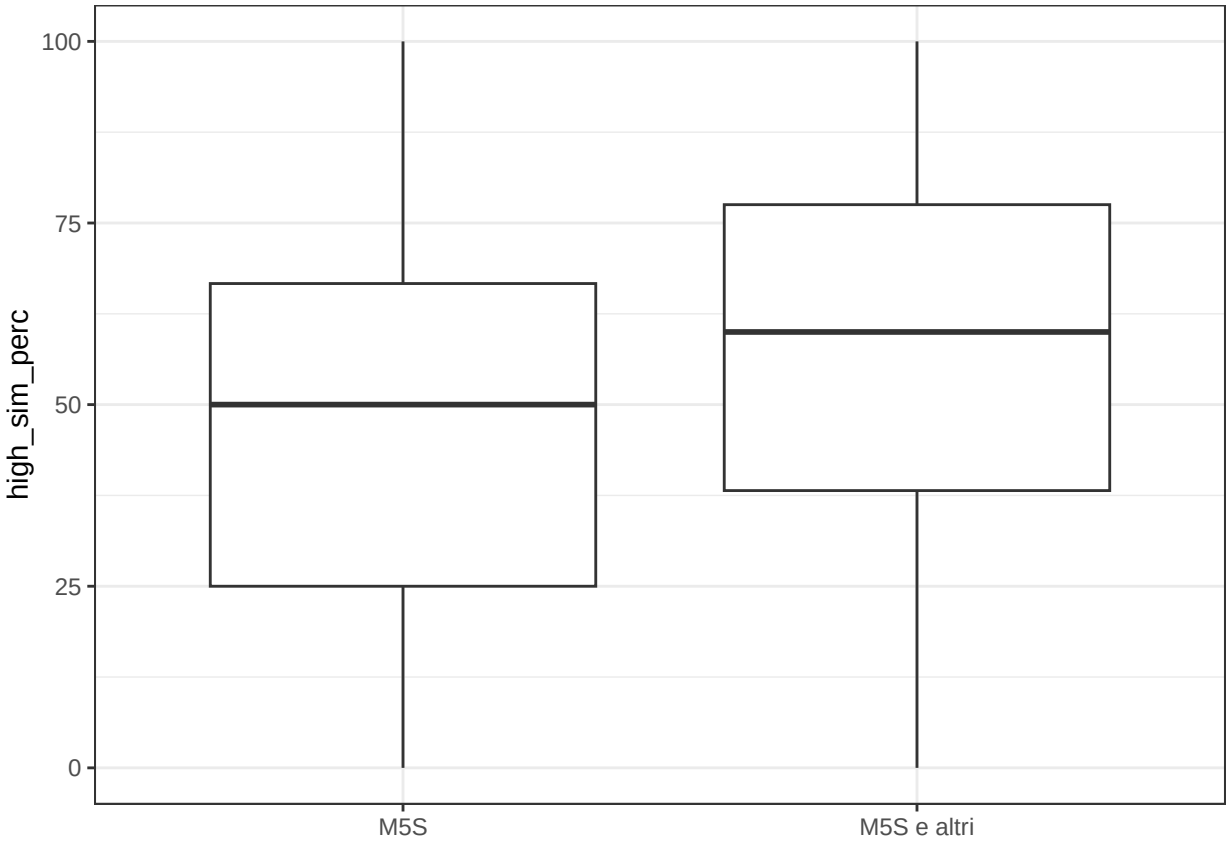


Figure 13: High similarity (%) by Partito

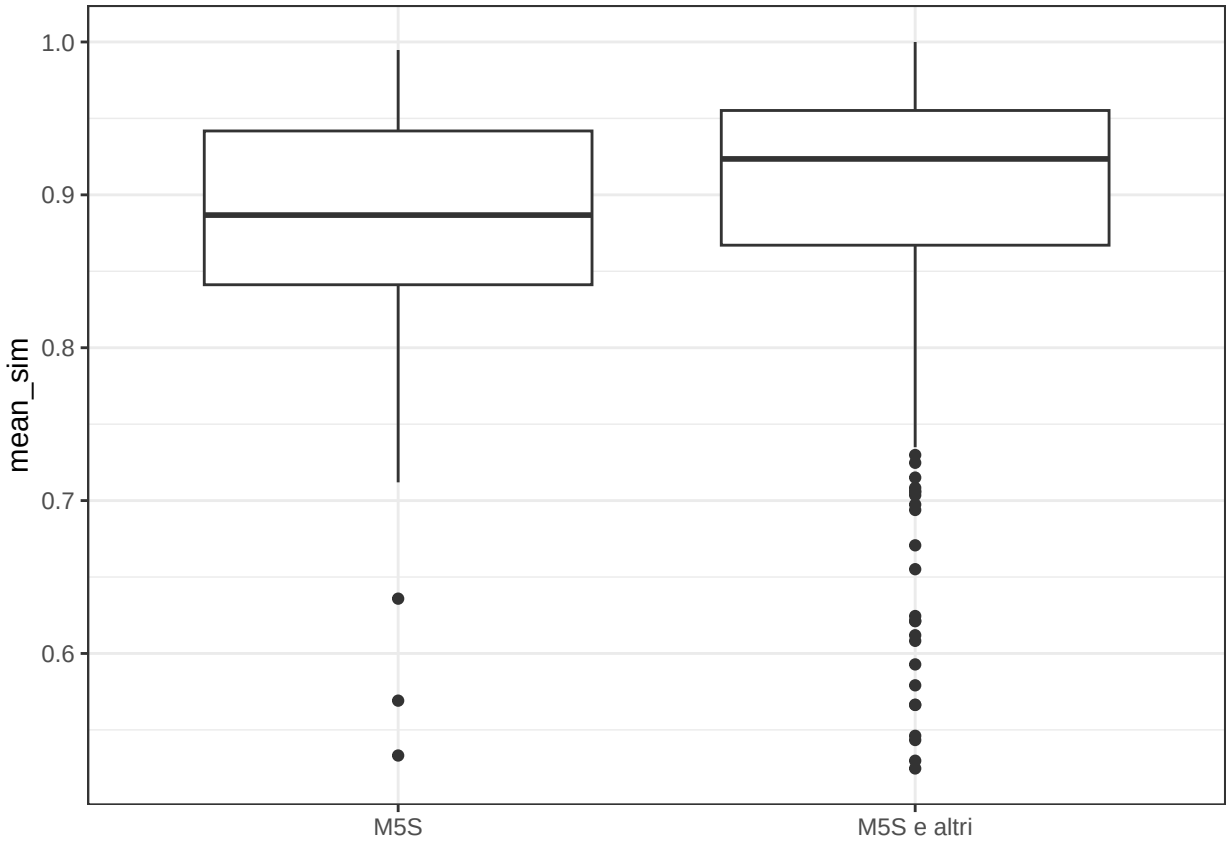


Figure 14: Mean similarity by Partito

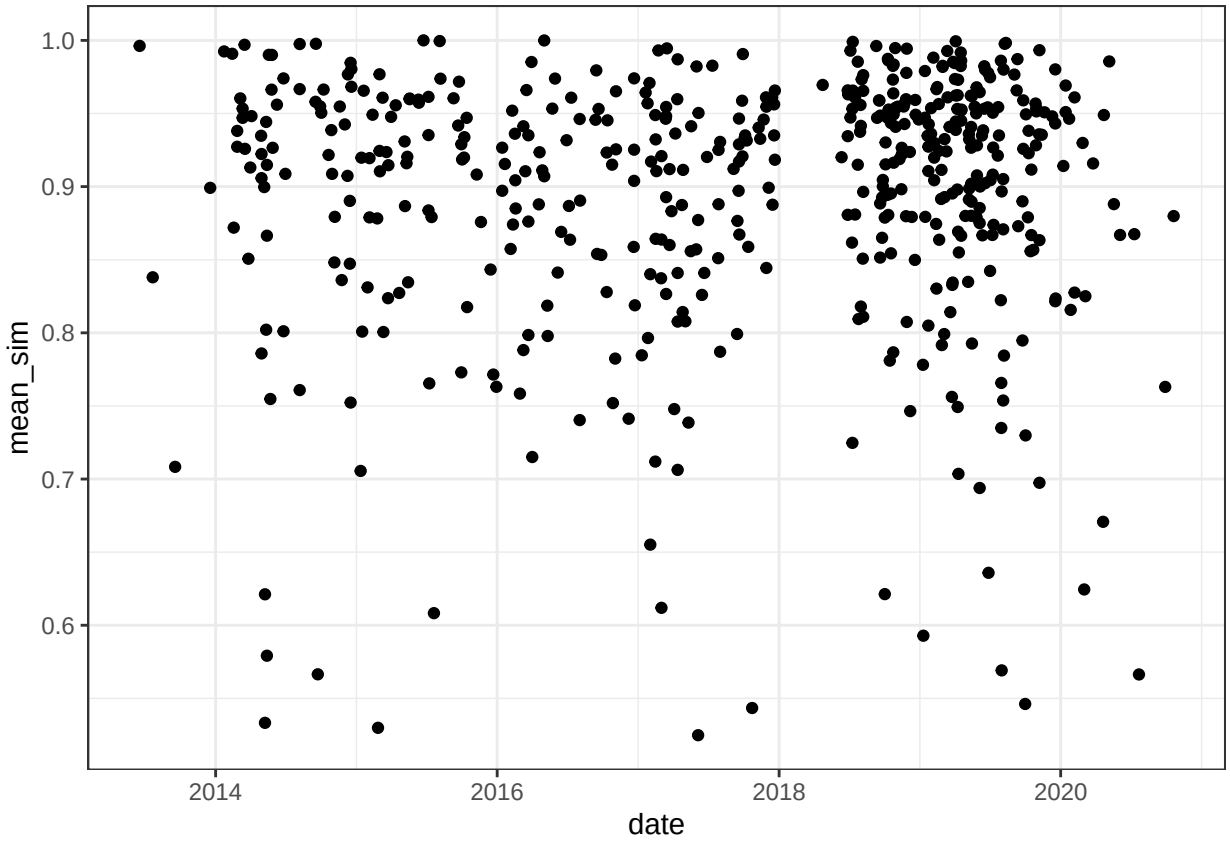


Figure 15: Mean similarity by time

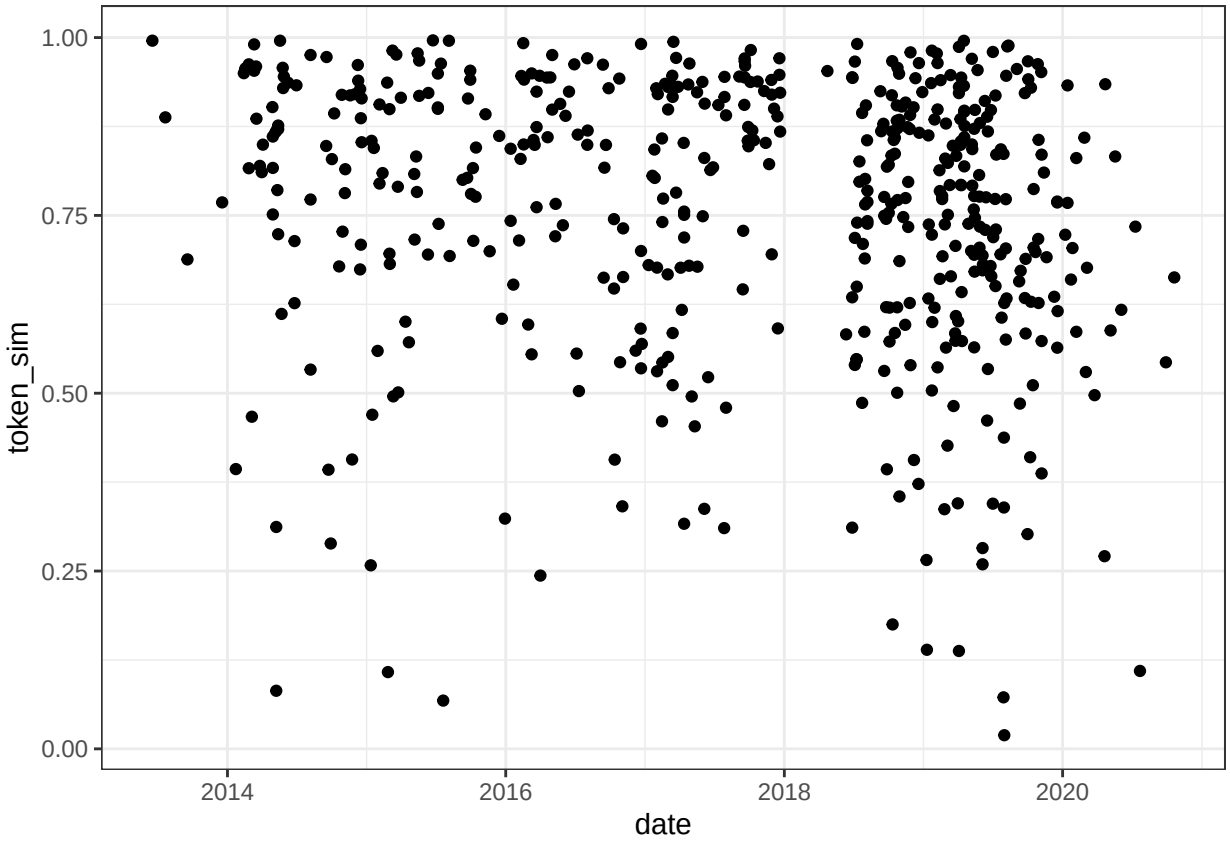


Figure 16: Token similarity by time

Table 3:

	Dependent variable:		
	mean_sim	bottom_5pct	bottom_10pct
	OLS	*logistic*	*logistic*
	(1)	(2)	(3)
Partito.nM5S only	−0.020*	0.167	0.548
	(0.012)	(0.572)	(0.386)
sentences	−0.00004	−0.002	0.003
	(0.0001)	(0.006)	(0.002)
'n commenti'	0.00002	−0.003	−0.00004
	(0.00003)	(0.004)	(0.001)
'n risposte firmatario'	−0.0004*	0.021*	0.015*
	(0.0003)	(0.011)	(0.009)
time_period.nXVIII leg. - Conte I	0.019**	−0.763	−0.227
	(0.009)	(0.562)	(0.352)
time_period.nXVIII leg. - Conte II	−0.011	0.444	0.386
	(0.014)	(0.633)	(0.475)
Constant	0.892***	−2.653***	−2.375***
	(0.007)	(0.483)	(0.276)
Observations	516	516	516
R^2	0.029		
Adjusted R^2	0.017		
Log Likelihood		−97.245	−163.730
Akaike Inf. Crit.		208.489	341.461
Residual Std. Error	0.088 (df = 509)		
F Statistic	2.515** (df = 6; 509)		
Note:			*p<0.1; **p<0.05; ***p<0.01

Full results

Below in the table, for each document the number of sentences in the DOCX and the number of sentences with a high similarity in the PDF, where the cosine similarity of at least one sentence pair (DOCX-PDF) was above .95.

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
1	65	0.97	0.95	46	70.77
3	60	0.87	0.83	10	16.67
4	8	0.64	0.62	0	0.00
5	8	0.84	0.87	2	25.00
7	2	0.99	0.99	2	100.00
9	166	0.96	0.92	91	54.82
10	7	1.00	0.98	6	85.71
11	28	0.83	0.76	4	14.29
12	5	0.97	0.96	4	80.00
13	14	0.96	0.95	10	71.43
14	1	0.98	0.98	1	100.00
15	64	0.92	0.86	28	43.75
16	36	0.97	0.92	22	61.11
18	64	0.99	0.99	60	93.75
19	5	0.97	0.96	4	80.00
20	11	0.94	0.84	4	36.36
21	32	0.98	0.93	21	65.62
22	39	0.98	0.95	30	76.92
23	49	0.98	0.91	35	71.43
24	4	0.98	0.96	3	75.00
25	12	0.97	0.95	10	83.33
26	42	0.94	0.90	20	47.62
27	73	0.99	0.97	61	83.56
28	24	0.96	0.93	13	54.17
29	6	0.95	0.94	3	50.00
31	20	0.96	0.93	13	65.00
32	5	0.79	0.77	1	20.00
33	3	0.94	0.96	1	33.33
35	112	0.70	0.71	10	8.93
36	3	1.00	1.00	3	100.00
37	2	0.94	0.94	1	50.00
38	90	0.90	0.89	39	43.33
39	15	0.93	0.92	4	26.67
40	27	0.96	0.91	16	59.26
41	17	0.87	0.85	5	29.41
42	9	0.80	0.81	0	0.00
43	25	0.73	0.74	2	8.00
44	24	1.00	1.00	24	100.00
45	23	0.99	0.83	15	65.22
46	4	0.88	0.83	0	0.00
47	78	0.92	0.89	32	41.03
49	9	0.99	0.96	6	66.67
50	12	0.96	0.90	6	50.00

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
52	8	0.97	0.95	6	75.00
53	3	0.99	0.98	3	100.00
54	39	0.82	0.79	3	7.69
55	4	1.00	0.87	3	75.00
58	4	0.85	0.80	2	50.00
59	6	0.96	0.95	3	50.00
60	7	0.98	0.92	6	85.71
61	24	0.98	0.95	15	62.50
62	15	0.97	0.93	10	66.67
63	6	0.95	0.96	3	50.00
64	9	0.94	0.92	4	44.44
65	4	0.94	0.94	1	25.00
66	78	0.98	0.93	58	74.36
67	2	0.75	0.75	1	50.00
68	11	0.81	0.81	0	0.00
69	13	1.00	1.00	13	100.00
70	9	0.95	0.93	5	55.56
71	69	0.99	0.98	62	89.86
72	9	0.99	0.94	8	88.89
74	14	0.80	0.72	0	0.00
75	61	0.94	0.91	28	45.90
76	154	0.99	0.97	115	74.68
77	36	0.96	0.90	20	55.56
78	18	0.91	0.87	7	38.89
79	30	0.93	0.90	12	40.00
80	14	0.95	0.87	7	50.00
81	5	0.95	0.75	3	60.00
82	11	0.95	0.90	6	54.55
83	11	0.96	0.87	6	54.55
84	5	0.96	0.96	4	80.00
85	4	0.71	0.69	0	0.00
86	6	0.97	0.97	4	66.67
87	7	0.96	0.90	6	85.71
88	5	0.99	0.96	3	60.00
89	1	0.99	0.99	1	100.00
90	3	0.98	0.88	2	66.67
91	2	0.62	0.62	1	50.00
92	16	0.91	0.88	6	37.50
93	16	1.00	1.00	16	100.00
94	11	0.95	0.88	5	45.45
95	4	0.97	0.82	3	75.00
96	26	0.98	0.94	18	69.23
97	13	0.97	0.95	10	76.92
98	67	0.96	0.91	39	58.21
99	35	0.97	0.92	25	71.43
100	155	0.97	0.92	94	60.65
101	77	0.97	0.95	52	67.53

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
103	152	1.00	0.99	148	97.37
104	15	0.99	0.91	12	80.00
105	3	0.89	0.89	0	0.00
106	30	0.98	0.93	21	70.00
107	13	0.93	0.91	4	30.77
108	10	0.94	0.87	5	50.00
109	6	0.81	0.84	2	33.33
110	9	0.90	0.86	2	22.22
111	3	0.99	0.98	3	100.00
112	7	0.93	0.79	3	42.86
113	3	0.95	0.94	2	66.67
114	12	0.89	0.82	1	8.33
115	2	0.54	0.54	1	50.00
116	107	0.92	0.84	38	35.51
117	79	0.97	0.94	55	69.62
118	12	0.88	0.89	5	41.67
119	845	0.76	0.75	129	15.27
120	15	0.76	0.78	3	20.00
121	2	0.86	0.86	0	0.00
122	12	0.93	0.86	5	41.67
123	20	0.97	0.90	11	55.00
124	19	0.98	0.96	13	68.42
125	12	0.90	0.89	3	25.00
127	16	0.98	0.97	14	87.50
128	13	0.95	0.88	7	53.85
129	17	0.96	0.95	13	76.47
130	15	0.95	0.88	8	53.33
131	29	1.00	0.99	28	96.55
132	3	0.95	0.84	2	66.67
133	5	0.97	0.91	3	60.00
134	4	0.92	0.85	2	50.00
135	3	0.94	0.92	1	33.33
136	23	0.98	0.97	18	78.26
137	172	0.97	0.95	120	69.77
138	3	1.00	0.99	3	100.00
139	25	1.00	1.00	25	100.00
140	21	0.93	0.89	8	38.10
141	132	0.93	0.91	55	41.67
145	17	1.00	0.98	15	88.24
146	10	0.98	0.96	9	90.00
147	14	1.00	0.99	13	92.86
148	18	1.00	1.00	18	100.00
149	31	0.44	0.52	2	6.45
150	28	0.95	0.88	14	50.00
151	9	0.99	0.95	7	77.78
152	6	0.97	0.82	4	66.67
153	4	0.98	0.98	4	100.00

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
155	19	0.89	0.85	3	15.79
156	2	0.84	0.84	0	0.00
157	2	0.77	0.77	1	50.00
158	2	0.88	0.88	0	0.00
159	27	0.99	0.96	19	70.37
160	6	0.92	0.85	3	50.00
161	10	0.99	0.91	8	80.00
162	9	1.00	1.00	9	100.00
163	1	0.99	0.99	1	100.00
164	14	0.85	0.85	2	14.29
165	14	0.93	0.91	4	28.57
166	9	0.95	0.88	3	33.33
167	25	0.99	0.99	25	100.00
168	1	0.88	0.88	0	0.00
169	2	0.92	0.92	1	50.00
170	4	0.75	0.75	2	50.00
171	4	0.68	0.59	0	0.00
172	11	0.97	0.97	8	72.73
173	35	0.98	0.96	27	77.14
174	11	1.00	0.99	11	100.00
175	11	0.94	0.88	5	45.45
176	18	0.98	0.94	14	77.78
177	22	0.96	0.87	13	59.09
178	49	0.96	0.95	29	59.18
179	9	0.91	0.91	1	11.11
180	10	0.99	0.99	10	100.00
181	5	0.86	0.80	0	0.00
182	8	0.72	0.75	1	12.50
183	107	0.81	0.80	28	26.17
184	42	1.00	0.93	36	85.71
185	19	0.92	0.88	6	31.58
186	11	0.98	0.92	7	63.64
188	129	0.79	0.81	33	25.58
189	5	0.97	0.96	3	60.00
190	20	0.94	0.88	10	50.00
191	4	0.81	0.78	0	0.00
192	2	0.96	0.96	2	100.00
193	12	0.95	0.94	5	41.67
194	21	1.00	1.00	21	100.00
195	7	0.93	0.93	3	42.86
196	10	0.97	0.95	6	60.00
197	5	0.97	0.85	3	60.00
199	8	0.99	0.98	7	87.50
200	1	0.84	0.84	0	0.00
201	2	0.70	0.70	0	0.00
202	22	0.73	0.73	0	0.00
203	3	0.98	0.89	2	66.67
204	6	1.00	0.96	5	83.33

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
205	1	0.57	0.57	0	0.00
206	10	0.92	0.89	2	20.00
207	54	1.00	0.99	53	98.15
208	45	0.96	0.88	25	55.56
209	7	0.97	0.93	6	85.71
210	35	0.87	0.88	11	31.43
211	11	0.90	0.87	1	9.09
212	10	0.88	0.79	1	10.00
213	10	0.96	0.96	7	70.00
214	36	0.95	0.90	21	58.33
215	38	0.97	0.93	21	55.26
216	43	1.00	0.96	35	81.40
217	3	0.70	0.70	0	0.00
220	26	0.98	0.95	21	80.77
221	20	0.99	0.97	16	80.00
222	22	0.85	0.78	2	9.09
223	13	0.98	0.97	12	92.31
224	7	0.99	0.95	4	57.14
225	12	0.96	0.94	7	58.33
226	21	0.99	0.96	15	71.43
227	13	0.94	0.88	6	46.15
228	1	0.61	0.61	0	0.00
229	7	0.97	0.91	6	85.71
230	40	1.00	0.99	38	95.00
231	20	0.90	0.84	6	30.00
232	13	1.00	0.95	10	76.92
235	9	0.99	0.91	5	55.56
238	27	1.00	0.99	26	96.30
239	90	0.79	0.76	5	5.56
240	6	0.98	0.93	5	83.33
241	42	0.98	0.93	31	73.81
242	186	0.95	0.92	95	51.08
244	8	0.93	0.93	3	37.50
245	9	0.97	0.97	6	66.67
247	8	0.93	0.92	3	37.50
249	14	0.82	0.83	3	21.43
250	45	0.87	0.83	7	15.56
252	11	0.97	0.93	7	63.64
253	19	0.99	0.95	15	78.95
254	37	0.99	0.98	35	94.59
255	17	0.98	0.95	12	70.59
256	10	0.95	0.93	4	40.00
257	8	0.98	0.93	5	62.50
258	18	0.81	0.76	3	16.67
259	2	0.80	0.80	0	0.00
260	52	0.96	0.93	27	51.92
261	11	0.96	0.95	9	81.82

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
262	20	0.89	0.85	3	15.00
263	17	0.95	0.90	8	47.06
264	8	0.95	0.94	5	62.50
265	1	0.97	0.97	1	100.00
266	1	0.98	0.98	1	100.00
267	11	1.00	0.99	10	90.91
268	4	0.98	0.98	4	100.00
269	3	0.98	0.97	2	66.67
270	22	0.98	0.95	17	77.27
271	7	0.97	0.96	5	71.43
272	33	0.95	0.90	14	42.42
273	4	0.95	0.87	2	50.00
275	43	0.84	0.82	4	9.30
277	2	0.78	0.78	1	50.00
278	8	0.89	0.89	2	25.00
279	1	0.99	0.99	1	100.00
280	27	0.99	0.96	19	70.37
281	73	1.00	1.00	73	100.00
282	10	0.99	0.91	7	70.00
283	90	0.96	0.91	48	53.33
284	10	0.99	0.99	10	100.00
285	30	1.00	1.00	29	96.67
286	4	0.97	0.92	3	75.00
287	15	0.96	0.94	10	66.67
288	9	0.93	0.77	4	44.44
289	11	0.97	0.89	8	72.73
290	14	0.98	0.92	11	78.57
291	2	0.79	0.79	1	50.00
292	2	0.95	0.95	1	50.00
293	10	0.94	0.90	4	40.00
295	24	0.98	0.97	20	83.33
296	48	0.97	0.95	32	66.67
297	21	0.96	0.93	13	61.90
299	30	0.95	0.91	15	50.00
300	20	1.00	1.00	20	100.00
301	6	0.88	0.88	1	16.67
302	13	0.94	0.93	5	38.46
303	4	0.70	0.71	0	0.00
304	34	0.93	0.90	13	38.24
305	230	0.95	0.90	119	51.74
306	34	0.98	0.93	24	70.59
307	2	0.80	0.80	0	0.00
308	4	0.93	0.83	2	50.00
309	7	0.98	0.97	5	71.43
310	7	0.95	0.90	4	57.14
313	32	0.93	0.88	14	43.75
314	26	0.74	0.76	1	3.85
315	24	0.96	0.86	13	54.17

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
317	48	1.00	0.93	37	77.08
318	145	0.97	0.94	90	62.07
319	3	0.89	0.91	1	33.33
320	14	0.99	0.98	12	85.71
321	30	0.92	0.88	8	26.67
322	35	0.94	0.89	17	48.57
323	35	0.97	0.95	28	80.00
324	8	0.87	0.88	1	12.50
325	16	0.85	0.86	4	25.00
326	12	0.89	0.88	2	16.67
328	34	0.98	0.92	24	70.59
329	26	0.95	0.89	11	42.31
330	12	0.98	0.97	10	83.33
331	6	0.99	0.97	5	83.33
332	1	0.96	0.96	1	100.00
333	25	1.00	0.96	21	84.00
334	17	0.98	0.95	12	70.59
335	32	0.95	0.88	16	50.00
336	8	0.99	0.86	6	75.00
337	34	0.98	0.96	28	82.35
338	48	0.96	0.91	28	58.33
339	24	0.93	0.87	10	41.67
340	5	0.88	0.87	2	40.00
341	9	0.98	0.93	5	55.56
342	13	0.98	0.96	11	84.62
343	7	0.97	0.83	4	57.14
344	17	0.99	0.97	16	94.12
345	16	0.93	0.86	7	43.75
346	16	1.00	0.98	14	87.50
347	11	0.97	0.95	8	72.73
348	2	0.85	0.85	0	0.00
349	10	0.80	0.81	0	0.00
350	1	0.89	0.89	0	0.00
351	6	0.96	0.95	4	66.67
352	41	0.99	0.97	33	80.49
353	10	0.87	0.80	1	10.00
354	54	0.94	0.89	26	48.15
355	28	0.80	0.81	3	10.71
356	19	0.99	0.97	14	73.68
357	3	0.94	0.94	1	33.33
358	24	0.98	0.97	20	83.33
359	2	0.98	0.98	2	100.00
361	7	0.81	0.79	1	14.29
363	10	0.90	0.84	4	40.00
364	2	0.80	0.80	1	50.00
365	8	0.99	0.99	8	100.00
366	11	0.99	0.97	10	90.91

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
367	15	0.95	0.94	7	46.67
369	4	0.73	0.73	2	50.00
370	20	0.97	0.96	17	85.00
372	1	0.99	0.99	1	100.00
373	1	0.93	0.93	0	0.00
375	19	0.86	0.78	5	26.32
377	2	0.90	0.90	0	0.00
378	78	0.88	0.85	19	24.36
379	46	0.96	0.92	27	58.70
380	8	0.98	0.97	7	87.50
382	16	0.98	0.95	13	81.25
383	10	0.92	0.83	4	40.00
384	3	0.59	0.67	0	0.00
385	12	0.99	0.95	10	83.33
386	3	0.99	0.93	2	66.67
387	30	0.98	0.95	23	76.67
388	14	0.86	0.86	3	21.43
389	96	0.91	0.88	31	32.29
390	122	0.83	0.84	44	36.07
391	9	0.98	0.94	6	66.67
392	9	0.48	0.57	1	11.11
393	12	0.80	0.82	1	8.33
394	9	0.86	0.87	2	22.22
395	9	0.98	0.97	7	77.78
396	9	0.95	0.93	5	55.56
397	60	0.94	0.91	25	41.67
398	390	0.96	0.93	226	57.95
400	5	0.93	0.81	2	40.00
401	5	0.89	0.91	2	40.00
402	8	0.94	0.85	3	37.50
403	19	0.99	0.94	16	84.21
404	77	0.59	0.62	1	1.30
405	51	0.53	0.53	0	0.00
407	38	0.94	0.92	16	42.11
408	14	0.98	0.94	9	64.29
409	10	0.99	0.98	10	100.00
410	43	0.94	0.89	20	46.51
411	27	0.98	0.96	20	74.07
412	3	0.97	0.86	2	66.67
413	61	0.95	0.90	30	49.18
414	4	0.88	0.79	1	25.00
415	7	0.96	0.96	5	71.43
416	53	0.98	0.92	33	62.26
417	7	0.95	0.91	4	57.14
419	3	0.98	0.98	3	100.00
421	21	0.97	0.85	11	52.38
422	37	0.97	0.86	20	54.05
424	4	0.82	0.75	2	50.00

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
425	18	0.95	0.84	9	50.00
426	43	0.93	0.88	17	39.53
427	25	0.90	0.80	7	28.00
428	9	0.97	0.96	7	77.78
429	7	0.96	0.94	4	57.14
430	15	0.96	0.92	10	66.67
431	8	0.99	0.93	6	75.00
432	16	0.99	0.93	14	87.50
433	9	0.99	0.92	8	88.89
434	6	0.96	0.96	5	83.33
435	8	0.97	0.92	4	50.00
436	7	0.99	0.95	6	85.71
437	6	0.93	0.81	2	33.33
438	17	0.91	0.83	8	47.06
439	25	0.99	0.97	22	88.00
441	11	0.97	0.89	7	63.64
442	3	0.95	0.95	1	33.33
443	11	0.76	0.80	3	27.27
444	17	0.99	0.95	13	76.47
446	9	0.85	0.86	1	11.11
447	40	0.98	0.94	27	67.50
448	16	0.98	0.95	13	81.25
449	1	0.86	0.86	0	0.00
450	8	0.97	0.94	5	62.50
451	17	0.97	0.95	14	82.35
452	55	0.53	0.53	0	0.00
453	1	1.00	1.00	1	100.00
454	16	0.73	0.71	3	18.75
455	13	0.96	0.91	7	53.85
456	3	0.89	0.82	0	0.00
457	5	0.98	0.95	4	80.00
458	3	0.99	0.99	3	100.00
459	19	0.92	0.83	6	31.58
460	7	0.96	0.95	4	57.14
461	8	0.98	0.98	7	87.50
462	7	0.97	0.95	5	71.43
463	5	0.99	0.96	4	80.00
464	38	0.97	0.95	26	68.42
465	7	1.00	0.99	7	100.00
466	11	0.99	0.92	7	63.64
468	48	0.98	0.97	37	77.08
469	2	0.93	0.93	0	0.00
470	12	0.62	0.64	0	0.00
471	48	0.91	0.86	16	33.33
472	33	0.99	0.95	25	75.76
473	25	0.98	0.95	15	60.00
474	10	0.67	0.66	0	0.00

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
475	5	0.95	0.90	3	60.00
476	8	0.92	0.89	3	37.50
477	9	0.99	0.98	8	88.89
478	19	0.97	0.96	11	57.89
479	4	0.93	0.92	2	50.00
480	5	0.52	0.61	1	20.00
481	22	0.80	0.76	1	4.55
482	68	0.97	0.96	50	73.53
483	5	0.89	0.83	0	0.00
484	6	0.97	0.95	4	66.67
485	3	0.98	0.96	2	66.67
486	59	0.91	0.82	19	32.20
487	7	0.95	0.83	4	57.14
488	4	0.97	0.96	3	75.00
489	13	0.81	0.77	2	15.38
490	17	1.00	0.98	15	88.24
491	19	0.97	0.95	13	68.42
492	94	1.00	0.99	88	93.62
493	9	0.80	0.74	0	0.00
494	2	0.72	0.72	0	0.00
495	15	0.98	0.91	9	60.00
497	12	0.97	0.91	9	75.00
498	10	0.99	0.95	8	80.00
499	2	0.75	0.75	0	0.00
500	2	0.97	0.97	2	100.00
502	23	1.00	0.99	21	91.30
503	12	0.93	0.90	5	41.67
504	10	0.93	0.87	4	40.00
505	17	0.98	0.94	11	64.71
506	45	1.00	0.97	40	88.89
507	1	0.95	0.95	1	100.00
508	122	0.96	0.92	65	53.28
510	62	1.00	0.98	56	90.32
511	37	0.98	0.97	24	64.86
512	97	0.58	0.57	1	1.03
513	14	0.94	0.92	4	28.57
514	33	0.90	0.87	11	33.33
515	7	0.95	0.95	4	57.14
516	6	0.96	0.95	4	66.67
517	46	1.00	0.95	35	76.09
518	6	0.94	0.87	2	33.33
519	31	0.98	0.94	23	74.19
520	25	1.00	0.98	20	80.00
521	20	0.94	0.86	9	45.00
522	26	0.93	0.90	11	42.31
523	11	0.97	0.92	8	72.73
524	16	0.95	0.89	9	56.25
525	5	0.98	0.96	3	60.00

(continued)

doc_id	sentences	median_sim	mean_sim	high_sim_n	high_sim_perc
526	3	0.95	0.95	2	66.67
527	11	0.92	0.79	5	45.45
528	20	0.98	0.97	17	85.00
529	15	0.93	0.92	5	33.33
530	6	0.99	0.98	5	83.33
531	21	0.98	0.95	17	80.95
532	15	0.97	0.94	11	73.33
533	13	0.96	0.95	8	61.54
534	12	0.97	0.95	10	83.33
535	2	0.93	0.93	0	0.00
536	7	0.47	0.55	0	0.00
538	17	0.97	0.91	10	58.82
539	13	0.94	0.92	6	46.15
540	48	0.98	0.94	35	72.92
541	14	0.98	0.96	9	64.29
542	1	0.95	0.95	1	100.00
543	4	0.87	0.82	0	0.00
544	48	0.94	0.92	23	47.92
545	30	0.94	0.90	13	43.33
547	11	0.99	0.94	8	72.73
548	2	0.99	0.99	2	100.00
550	21	0.95	0.94	12	57.14
552	6	0.93	0.89	2	33.33
553	4	0.97	0.95	3	75.00
554	24	0.97	0.95	18	75.00
555	38	0.68	0.71	7	18.42
556	4	1.00	0.87	3	75.00
557	32	0.97	0.96	23	71.88
558	48	0.95	0.92	25	52.08
559	63	0.98	0.93	41	65.08
560	77	0.98	0.94	56	72.73
561	13	0.99	0.90	9	69.23
562	45	0.99	0.93	33	73.33
563	42	0.75	0.74	3	7.14
564	18	0.87	0.82	6	33.33
565	12	0.98	0.94	10	83.33
566	12	0.85	0.80	2	16.67
567	30	0.96	0.93	16	53.33
568	13	0.96	0.96	9	69.23
570	4	0.98	0.97	4	100.00
571	10	0.98	0.92	7	70.00
572	104	0.97	0.94	72	69.23
574	170	0.96	0.93	102	60.00
575	11	0.97	0.95	7	63.64
576	9	0.52	0.58	1	11.11
579	23	0.92	0.89	3	13.04

Figure 1

Figure 2

References

- Benoit, K., & Obeng, A. (2023). readtext: Import and handling for plain and formatted text files [Manual]. https://CRAN.R-project.org/package=readtext
- Edoardo Federici. (2022). Sentence-bert-base, sentence-transformer for Italian. *Hugging Face*. https://doi.org/10.57967/hf/0112
- Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *JOSS*, 1(3). https://doi.org/10.21105/joss.00037
- Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084