



Semantic Textual Similarity of Courses based on Text Embeddings

Olivera Kitanović, Aleksandra Tomašević, Mihailo Škorić, Ranka Stanković and Ljiljana Kolonja Serbia, University of Belgrade: Faculty of Mining and Geology, Serbia





- **RQ1.** Can a text vectorization approach, based on the course descriptions, be useful to identify courses with high semantic similarity?
- **RQ2.** To what extent can similarities between professor's specialty be calculated based on course vectors?
- **RQ3.** How does course similarity detection differ when using vectorization based on the course description compared to using the course title and description?

Methodology

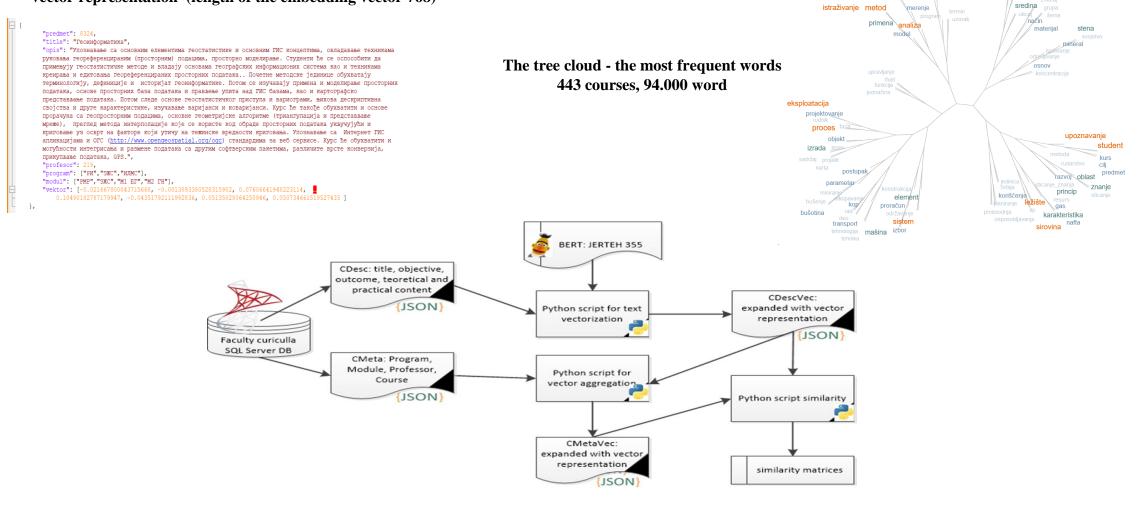


klasifikacija pojam

vrsta osobina

podatak

SQL data=>JSON file with course metadata, description and vector representation (length of the embedding vector 768)



Results



course similarity matrix 443x443

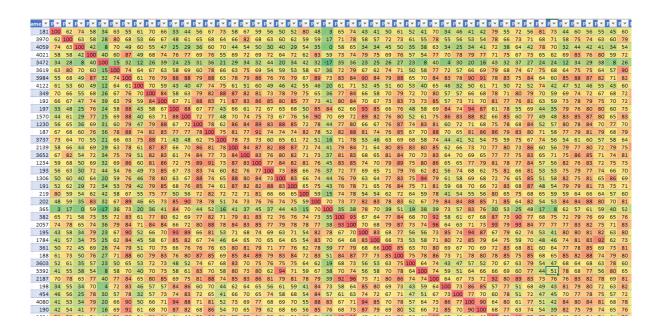
			_			"PCA"	~ .		~ ·	"PCA"	<u>.</u> .	ubethic		ibetuie i	111	- Anno 1	"PCA"	<u></u>	ubettine	•	in between	"PCA"		"PCAT	. .	repence		PCAILCI
Name 🕆	N	lin 📑	1	Max	-	8000	-	8001	-	8002	-	8003	- 8	3004 📑	80	005 👻	8006	-	8007	-	8008 🕆	8009	-	8010	-	8011	- 2	8012 👻
8000	C).8535	5	0.987	33		1	0.940	97	0.941	54	0.9402	23	0.9126	7 C	.98733	0.924	804	0.9240	07	0.9376	0.94	717	0.941	.94	0.925	01	0.91067
8001	C	0.8834	2	0.978	314	0.940	97		1	0.953	858	0.9575	51	0.9280	8 0	.94025	0.961	.83	0.9562	28	0.95729	0.95	132	0.960	74	0.91	48	0.94327
8002	C	0.8715	1	0.990	944	0.941	54	0.953	858		1	0.9527	78	0.9173	9 0	.95072	0.938	891	0.951	۲7	0.94746	0.93	765	0.953	27	0.890	06	0.92863
8003	C	0.91314	4	0.992	89	0.940	23	0.957	751	0.952	278		1	0.940	6 0	.93526	0.962	236	0.9668	37	0.95085	0.96	512	0.971	.13	0.941	46	0.97005
8004	C	0.8769	7	0.985	605	0.912	67	0.928	808	0.917	' 39	0.940	06		1 0	.91497	0.915	92	0.924	15	0.92766	0.94	937	0.943	78	0.921	77	0.91727
8005	C).8553	7	0.987	33	0.987	33	0.940)25	0.950)72	0.9352	26	0.9149	7	1	0.925	555	0.9317	79	0.94694	0.94	603	0.942	69	0.90	97	0.9117
8006	C).89684	4	0.984	53	0.924	08	0.961	183	0.938	891	0.9623	36	0.9159	2 0	.92555		1	0.957	72	0.94205	0.93	911	0.956	32	0.896	84	0.96467
8007	C	0.8982	2	0.979	49	0.924	07	0.956	528	0.95	517	0.9668	37	0.924	5 C	.93179	0.95	572		1	0.97181	0.95	471	0.972	02	0.898	22	0.97538
8008	C	.8986	5	0.979	000	0.93	76	0.957	29	0.947	46	0.9508	35	0.9276	6 0	.94694	0.942	205	0.9718	31	1	0.9	639	0.956	93	0.906	72	0.94882
8009	C	0.9005	7	0.977	'92	0.947	17	0.951	132	0.937	65	0.9651	12	0.9493	7 C	.94603	0.939	911	0.9547	1	0.9639		1	0.967	27	0.945	21	0.94379
8010	C	0.9115	8	0.988	325	0.941	94	0.960)74	0.953	327	0.9711	13	0.9437	8 C	.94269	0.956	532	0.9720)2	0.95693	0.96	727		1	0.933	94	0.96314
8011	C	0.8606	5	0.955	22	0.925	01	0.91	48	0.890	006	0.9414	16	0.9217	7	0.9097	0.896	684	0.8982	22	0.90672	0.94	521	0.933	94		1	0.90495
8012	0	0.8948	7	0.988	89	0.910	67	0.943	327	0.928	863	0.9700)5	0.9172	7	0.9117	0.964	67	0.9753	88	0.94882	0.94	379	0.963	14	0.904	95	1

course similarity matrix with normalized values 0-100

С	E	F	G	Η	I	J	K	L	Μ	N	0	Ρ	Q	R	S	Т	U	V	W	Х	Y	Ζ	AA	AB
naziv	ID	Min	Max	V8000	/8001	/8002	/8003	/8004	V8005	/8006	/8007	/8008	/8009	/8010	/8012	/8013	/8014	/8015	/8017	V8018	/8019	/8020	V8022	V8023
Математика 1	8000	21	92	100	64	65	64	47	92	54	54	62	68	65	46	55	46	66	68	60	63	51	61	56
Хемија 8	8001	43	87	64	100	72	74	57	64	77	74	74	71	76	66	75	67	75	78	68	81	53	75	72
Физика 1	8002	36	94	65	72	100	72	50	70	63	71	68	62	72	57	60	60	74	85	63	72	48	74	66
Општа геологија	8003	48	96	64	74	72	100	64	61	77	80	70	79	83	82	85	75	83	79	82	81	64	75	76
Енглески језик 1	8004	27	91	47	57	50	64	100	49	49	55	56	70	66	50	55	51	66	58	57	65	91	61	59
Математика 2	8005	22	92	92	64	70	61	49	100	55	59	68	67	65	47	53	48	69	75	57	63	50	67	58
Минералогија	8006	41	91	54	77	63	77	49	55	100	74	65	63	74	79	85	77	78	69	72	78	48	62	64
Геоморфологија	8007	46	88	54	74	71	80	55	59	74	100	83	73	83	85	81	79	88	83	76	76	50	80	74
Нацртна геометриј	8008	45	87	62	74	68	70	56	68	65	83	100	78	74	69	68	62	86	85	71	67	54	85	78
Рачунарство у хидр	8009	44	87	68	71	62	79	70	67	63	73	78	100	80	66	70	64	81	78	79	75	69	81	80
Општа хидрологија	8010	47	93	65	76	72	83	66	65	74	83	74	80	100	78	79	88	84	85	83	90	62	80	82
Историјска геологи	8012	37	91	46	66	57	82	50	47	79	85	69	66	78	100	87	79	83	69	78	72	46	65	68
Петрологија 8	8013	44	93	55	75	60	85	55	53	85	81	68	70	79	87	100	81	82	71	80	79	51	67	72
	8014	30	00	10	07	00	75		40	77	70	00	0.4	00	70	04	100	70	75	70	07	45	60	74

- E1: title and course description
 - 29 courses > 95% of similarities
- E2: just description.
 - 30 courses > 95%

professor similarity matrix with normalized values 0-100



Evaluation



	Professor ID	Course ID	Symilar1	Eval1	Symilar2	Eval2	Symilar3	Eval3	Symilar4	Eval4	Symilar5	Eval5
1	540	8382	[8597, 90]	90	[8535, 81]		[8457, 78]		[8680, 76]		[8526, 76]	
2	540	8597	[8382, 90]	90	[8535, 81]		[8538, 78]		[8680, 78]		[8145, 78]	
3	4060	8322	[8323, 99]	99	[8139, 94]		[8745, 92]		[8376, 89]		[8338, 89]	
4	388	8513	[8483, 91]		[8546, 89]	5	[8471, 88]	3	[8514, 88]	3	[8201, 88]	1
5	388	8546	[8483, 90]		[8513, 89]	5	[8514, 89]		[8491, 88]		[8399, 88]	
6	168	8589	[8655,100]	100	[8463, 90]		[8525, 88]		[8587, 86]		[8542,86]	
7	168	8655	[8589,100]	100	[8463, 89]		[8525, 87]		[8587, 86]		[8542,86]	
8	1784	8768	[8766, 99]	99	[8769,99]	99	[8767, 98]	99	[8722, 86]		[8729, 84]	
9	1600	8652	[8510, 94]	95	[8509, 89]		[8649, 87]		[8625, 87]		[8288, 86]	

Rows 4 and 5 => calculated similarity <> manual evaluation

Rows 6 and 7 => complete match

Row 3 => "metallic" <> "non-metallic".

Row 8 => "flora"/"fauna"/"animals", "organism"/"microorganism"

Row 9 => same general topics, different wording

Thank you for your attention!

olivera.kitanovic@rgf.bg.ac.rs

https://tesla.rgf.bg.ac.rs/

A research conducted with the support of the Science Fund of the Republic of Serbia, project number: 7276.

Text Embeddings -Serbian Language Applications –

TESLA

ТЕСЛА



