АНАЛИЗ КОНКУРЕНТНЫХ СИСТЕМ

JINR MLiT

Программный инструмент рубрикации научных публикаций на основе современных нейросетевых технологий

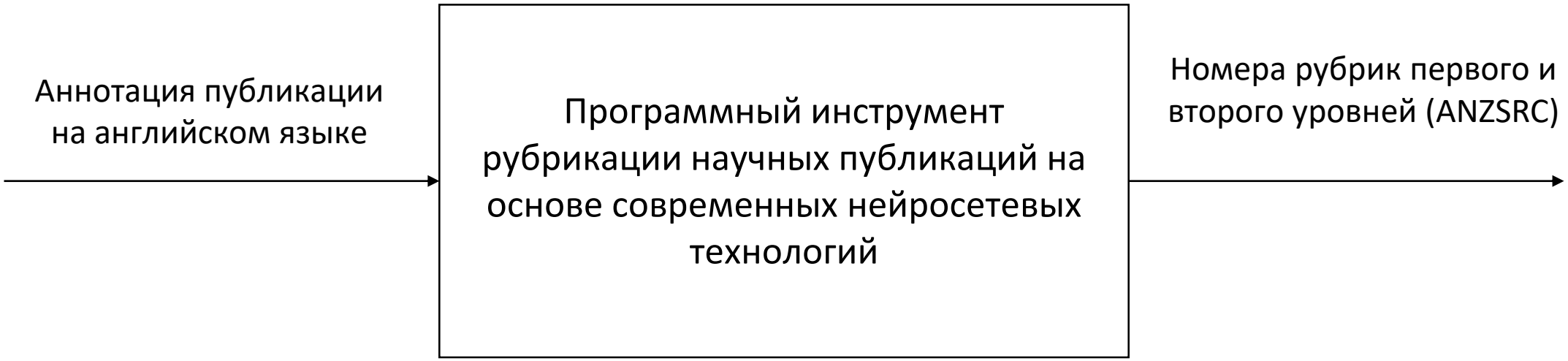
Студент: Патрушев Константин Алексеевич

Научный руководитель: Антонов Евгений Вячеславович

Дубна, 2024 г.

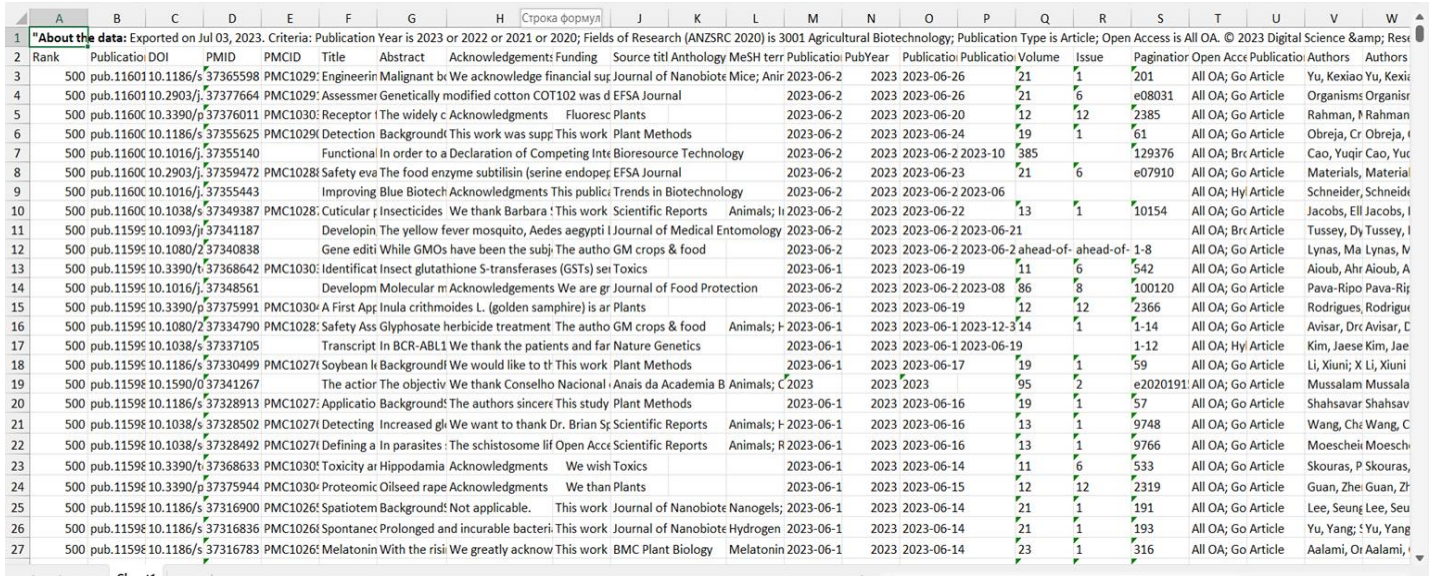
Цель работы

Разработка программного инструмента, осуществляющего рубрикацию научных публикаций по двум уровням рубрик (ANZSRC)



Описание исходных данных

- Стандарт рубрик - Australian and New Zealand Standard Research Classification
- 350 000 аннотаций выгружено из базы данных dimensions
- Для каждого текста аннотации представлены рубрики как 1-го (22 шт), так и 2-го уровней (171 шт)
- Тексты аннотаций на разных языках



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	**"About the data:** Exported on Jul 03, 2023. Criteria: Publication Year is 2023 or 2022 or 2021 or 2020; Fields of Research (ANZSRC 2020) is 3001 Agricultural Biotechnology; Publication Type is Article; Open Access is All OA. © 2023 Digital Science & Rese																						
2	Rank	Publicatio	DOI	PMID	PMCID	Title	Abstract	Acknowledgements	Funding	Source titl	Anthology	MeSH terr	Publicatio	PubYear	Publicatio	Publicatio	Volume	Issue	Paginatior	Open Acce	Publicatio	Authors	Authors
3	500	pub.1160	10.1186/s	37365598	PMC1029	Engineerin	Malignant b	We acknowledge	financial sup	Journal of Nanobiote		Mice; Anir	2023-06-2	2023	2023-06-26		21	1	201	All OA; Go	Article	Yu, Kexiao	Yu, Kexi
4	500	pub.1160	10.2903/j.	37377664	PMC1029	Assessmer	Genetically modified cotton COT102 was d			EFSA Journal			2023-06-2	2023	2023-06-26		21	6	e08031	All OA; Go	Article	Organisms	Organis
5	500	pub.1160	10.3390/p	37376011	PMC1030	Receptor i	The widely c	Acknowledgments	Fluoresc	Plants			2023-06-2	2023	2023-06-20		12	12	2385	All OA; Go	Article	Rahman, M	Rahman
6	500	pub.1160	10.1186/s	37355625	PMC1029	Detection	Background	This work was supp		This work		Plant Methods	2023-06-2	2023	2023-06-24		19	1	61	All OA; Go	Article	Obreja, Cr	Obreja,
7	500	pub.1160	10.1016/j.	37355140		Functional	In order to a	Declaration of Competing Inte		Bioresource Technology			2023-06-2	2023	2023-06-2	2023-10	385		129376	All OA; Br	Article	Cao, Yuqir	Cao, Yu
8	500	pub.1160	10.2903/j.	37359472	PMC1028	Safety eva	The food enzyme subtilisin (serine endopep			EFSA Journal			2023-06-2	2023	2023-06-23		21	6	e07910	All OA; Go	Article	Materials,	Materia
9	500	pub.1160	10.1016/j.	37355443		Improving	Blue Biotech	Acknowledgments	This public	Trends in Biotechnology			2023-06-2	2023	2023-06-2	2023-06				All OA; Hy	Article	Schneider,	Schneid
10	500	pub.1160	10.1038/s	37349387	PMC1028	Cuticular p	Insecticides	We thank Barbara S	This work	Scientific Reports		Animals; Ir	2023-06-2	2023	2023-06-22		13	1	10154	All OA; Go	Article	Jacobs, Ell	Jacobs,
11	500	pub.1159	10.1093/jr	37341187		Developin	The yellow fever mosquito, Aedes aegypti l			Journal of Medical Entomology			2023-06-2	2023	2023-06-2	2023-06-21				All OA; Br	Article	Tussey, Dy	Tussey,
12	500	pub.1159	10.1080/2	37340838		Gene editi	While GMOs have been the subj	The autho		GM crops & food			2023-06-2	2023	2023-06-2	2023-06-2	ahead-of-	ahead-of-	1-8	All OA; Go	Article	Lynas, Ma	Lynas, M
13	500	pub.1159	10.3390/t	37368642	PMC1030	Identificat	Insect glutathione S-transferases (GSTs) se			Toxics			2023-06-1	2023	2023-06-19		11	6	542	All OA; Go	Article	Aioub, Ahr	Aioub, A
14	500	pub.1159	10.1016/j.	37348561		Developm	Molecular m	Acknowledgements	We are gr	Journal of Food Protection			2023-06-2	2023	2023-06-2	2023-08	86	8	100120	All OA; Go	Article	Pava-Ripo	Pava-Ri
15	500	pub.1159	10.3390/p	37375991	PMC1030	A First App	Inula crithmoides L. (golden samphire) is ar			Plants			2023-06-1	2023	2023-06-19		12	12	2366	All OA; Go	Article	Rodrigues,	Rodrigu
16	500	pub.1159	10.1080/2	37334790	PMC1028	Safety Ass	Glyphosate herbicide treatment	The autho		GM crops & food		Animals; H	2023-06-1	2023	2023-06-1	2023-12-3	14	1	1-14	All OA; Go	Article	Avisar, Dr	Avisar, D
17	500	pub.1159	10.1038/s	37337105		Transcript	In BCR-ABL1	We thank the patients and far		Nature Genetics			2023-06-1	2023	2023-06-1	2023-06-19			1-12	All OA; Hy	Article	Kim, Jaese	Kim, Jae
18	500	pub.1159	10.1186/s	37330499	PMC1027	Soybean l	Background	We would like to th	This work	Plant Methods			2023-06-1	2023	2023-06-17		19	1	59	All OA; Go	Article	Li, Xiuni; X	Li, Xiuni
19	500	pub.1159	10.1590/0	37341267		The actior	The objectiv	We thank Conselho Nacional		Anais da Academia B		Animals; C	2023		2023	2023	95	2	e2020191	All OA; Go	Article	Mussalam	Mussala
20	500	pub.1159	10.1186/s	37328913	PMC1027	Applicatio	Background	The authors sincere	This study	Plant Methods			2023-06-1	2023	2023-06-16		19	1	57	All OA; Go	Article	Shahsavar	Shahsav
21	500	pub.1159	10.1038/s	37328502	PMC1027	Detecting	Increased gl	We want to thank Dr. Brian Sp		Scientific Reports		Animals; H	2023-06-1	2023	2023-06-16		13	1	9748	All OA; Go	Article	Wang, Cha	Wang, C
22	500	pub.1159	10.1038/s	37328492	PMC1027	Defining a	In parasites	The schistosome lif	Open Acce	Scientific Reports		Animals; R	2023-06-1	2023	2023-06-16		13	1	9766	All OA; Go	Article	Moescheir	Moesch
23	500	pub.1159	10.3390/t	37368633	PMC1030	Toxicity ar	Hippodamia	Acknowledgments	We wish	Toxics			2023-06-1	2023	2023-06-14		11	6	533	All OA; Go	Article	Skouras, P	Skouras,
24	500	pub.1159	10.3390/p	37375944	PMC1030	Proteomic	Oilseed rape	Acknowledgments	We than	Plants			2023-06-1	2023	2023-06-15		12	12	2319	All OA; Go	Article	Guan, Zhe	Guan, Z
25	500	pub.1159	10.1186/s	37316900	PMC1026	Spatiotem	Background	Not applicable.	This work	Journal of Nanobiote		Nanogels;	2023-06-1	2023	2023-06-14		21	1	191	All OA; Go	Article	Lee, Seung	Lee, Seu
26	500	pub.1159	10.1186/s	37316836	PMC1026	Spontanec	Prolonged and incurable bacteri		This work	Journal of Nanobiote		Hydrogen	2023-06-1	2023	2023-06-14		21	1	193	All OA; Go	Article	Yu, Yang; S	Yu, Yang
27	500	pub.1159	10.1186/s	37316783	PMC1026	Melatonin	With the risi	We greatly acknow	This work	BMC Plant Biology		Melatonin	2023-06-1	2023	2023-06-14		23	1	316	All OA; Go	Article	Aalami, Or	Aalami,

Пример выгруженного xlsx-файла с данными

Процесс разработки рубрикатора

Получение векторных представлений текстов аннотаций

allenai/longformer-base-4096 Размерность получаемого эмбеддинга: 768	⟹	MLP (скрытый слой: 256 нейронов, функция активации: ReLU)[1, 2]
Построение собственного словаря токенов модель fasttext для векторизации токенов	⟹	CNN (3 сверточных слоя, батч-нормализация после каждого сверточного слоя, слой классификации)
Токенизатор и векторизатор модели BERT-BASE-UNCASED	⟹	Предобученная модель BERT-BASE-UNCASED

1. Amjad, M., Gelbukh, A., Voronkov, I., & Saenko, A. (2019). Comparison of Text Classification Methods using Deep Learning Neural Networks

2.M, K., & Prabhavathy, P. (2023). Multi-Label Text Classification using Machine Learning Techniques

Результаты классификации для различных моделей

Для оценки качества классификации производится подсчет метрики **Accuracy** на тестовой (OOS) выборке

$$\textbf{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

Рубрикатор 1-го уровня

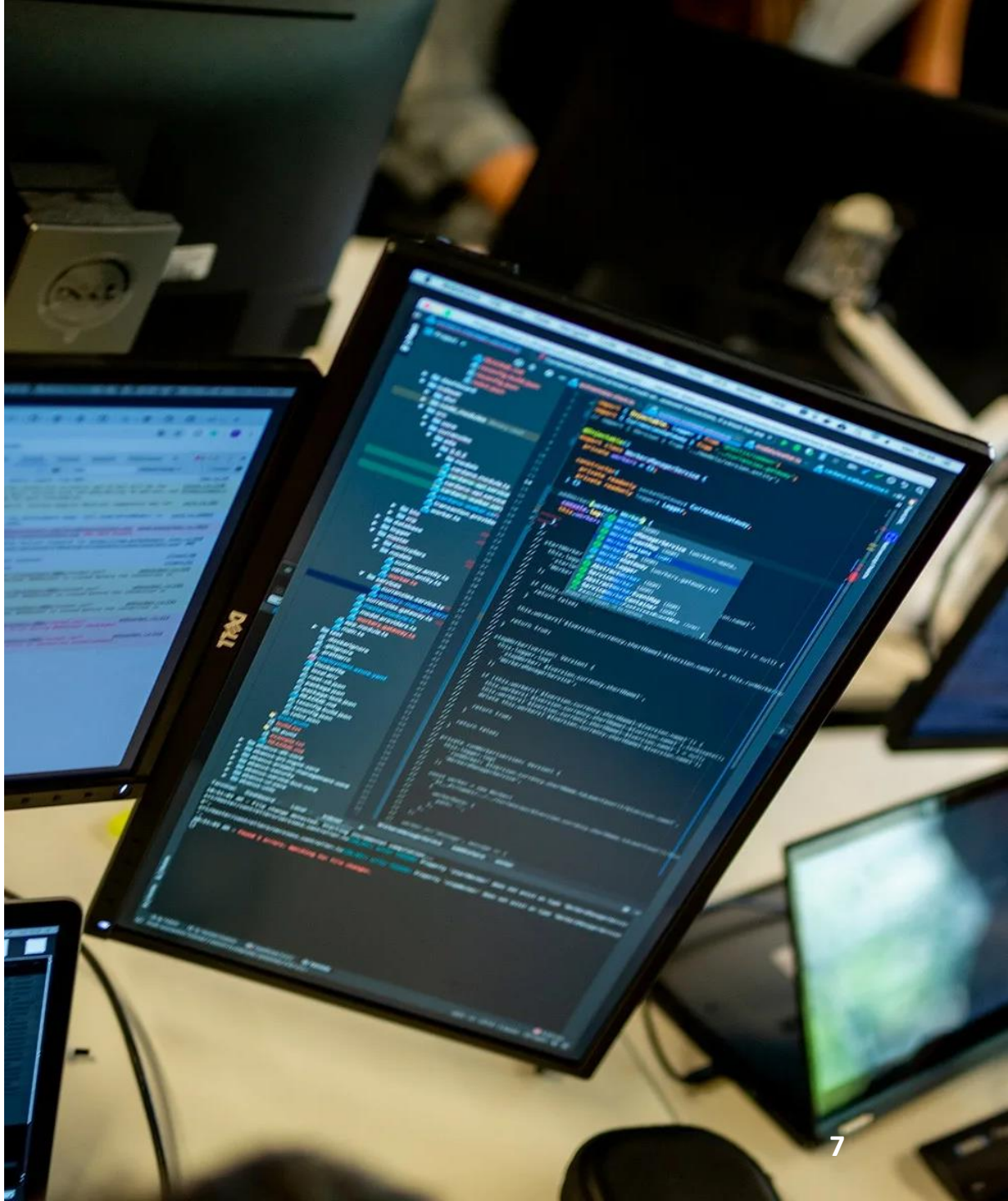
1. MLP
    1.1 Accuracy Test: 0.659
2. CNN
    2.1 Accuracy Test: 0.687
3. BERT
    3.2 Accuracy Test: 0.84

Рубрикатор 2-го уровня

1. MLP
    1.1 Accuracy Test: 0.1103
2. CNN
    2.1 Accuracy Test: 0.1342
3. BERT
    3.1 Accuracy Test: 0.5

Результаты и выводы

- Разработан программный инструмент рубрикации научных публикаций по двум уровням
- Проанализированы различные нейросетевые архитектуры для решения поставленной задачи
- Оценено качество работы моделей на основе различных нейросетевых архитектур

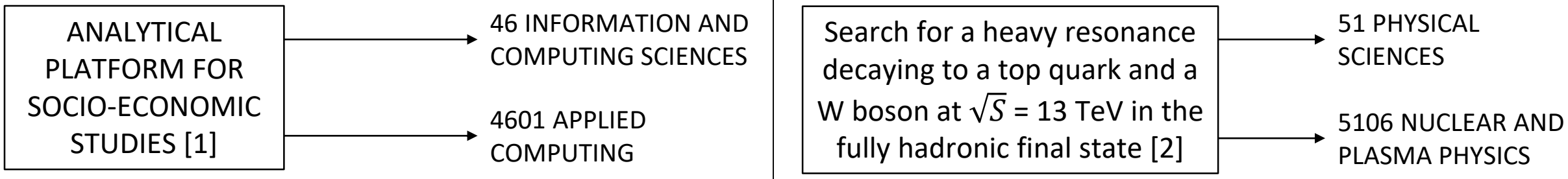


Результаты и выводы

- Рубрикаторы 2-го уровня имеют низкое качество

Результаты и выводы

- Демонстрация работы программного инструмента рубрикации научных публикаций



1. Belov, S., Ilina, A. V., Javadzade, J., Kadochnikov, I., Korenkov, V., Pelevanyuk, I., Tarabrin, V., Zrelov, P., & Semenov, R. (2021). ANALYTICAL PLATFORM FOR SOCIO-ECONOMIC STUDIES. 9th International Conference “Distributed Computing and Grid Technologies in Science and Education.”

2. Sirunyan, A. M., Tumasyan, A., Adam, W., Bergauer, T., Shmatov, S. V. (2021). Search for a heavy resonance decaying to a top quark and a W boson at sqrt{s} = 13 TeV in the fully hadronic final state. ˜the œJournal of High Energy Physics/˜the Journal of High Energy Physics, 2021(12).

Дальнейшие планы

- Улучшение качества рубрикации 2-го уровня
- Разработка API для использования инструмента

Спасибо за внимание!