

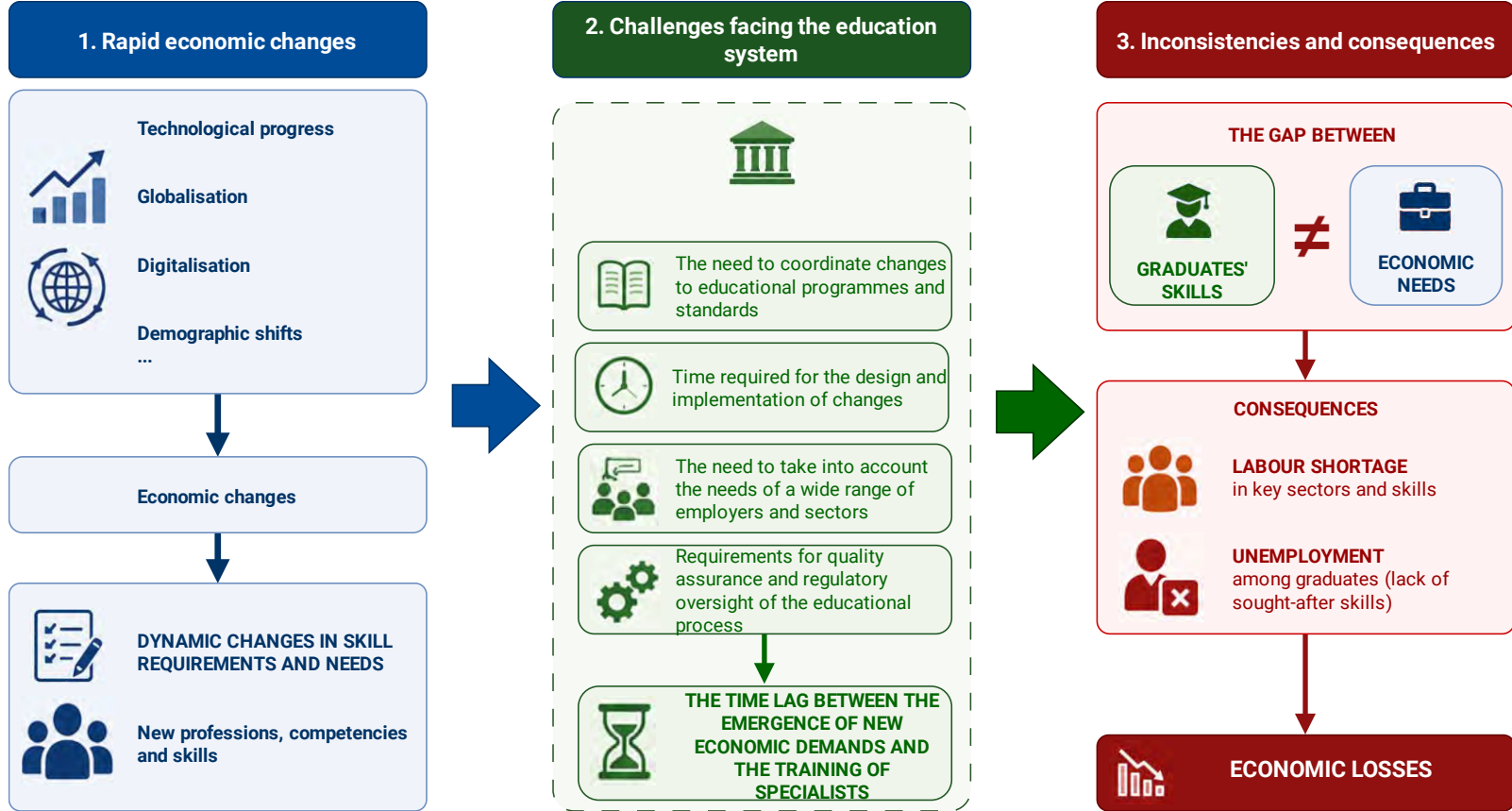
Сравнительный анализ методов объединения семантических моделей в задаче сопоставления вакансий рынка труда и образовательных компетенций

Comparative analysis of methods for combining semantic models in the task of comparing labor market vacancies and educational competencies

Ilina A.V.*, Starchenko Ju.B., Tarabrin V.A., Zrellov P.V.

*Laboratory of Information Technologies named after M.G. Meshcheryakov
Joint Institute for Nuclear Research (JINR)*

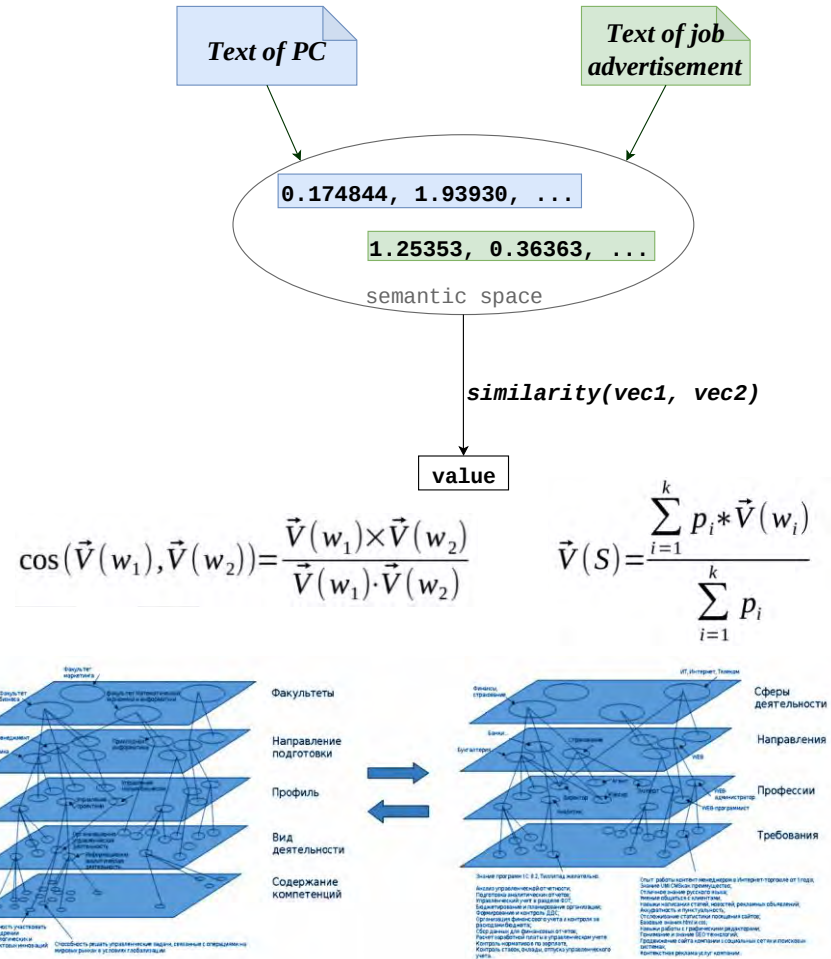
* annailina@jinr.ru



Studying these mechanisms is essential for developing effective educational policies and labor demand forecasting tools.

Previous results. The idea

The technique developed by the authors¹⁻⁵ is based on translating texts into semantic space using embedding models, which allows a researcher using vector algebra methods to find the distances between the resulting vectors, thereby determining the value of semantic similarity of texts.



¹ Monitoring of labor market needs in university graduates based on data-intensive analytics / P. V. Zrellov, V. V. Korenkov, N. A. Kutovsky, A. Sh. Petrosyan [et al.] // Proceedings of XVIII International Conf. DAMDID/RCDL 2016 "Analytics and data management in data intensive domains", Ershovo, October 11-14, 2016. -Moscow: Federal Research Center "Informatics and Control" of the Russian Academy of Sciences, 2016. -С. 124-131.

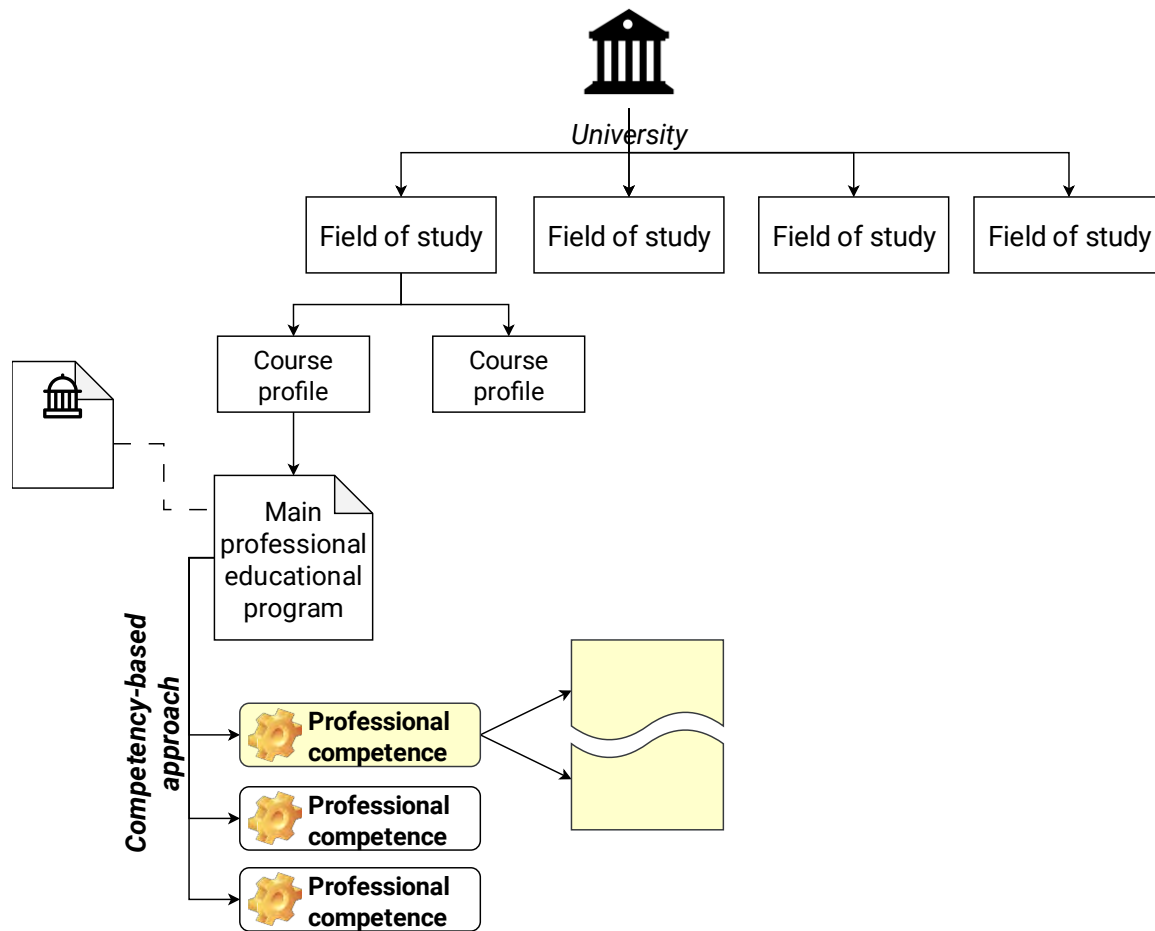
² Monitoring the compliance of professional education with the needs of the labor market / S.D.Valentey, P.V.Zrellov, V.V.Korenkov, S.D.Belov[et al.] // Social Sciences and Modernity. 2018. - No3. - pp. 5-16.

³ Methods and algorithms of the analytical platform for analyzing the labor market and the compliance of the higher education system with market needs / S. Belov [et al.] // Proceedings of Science, 2022. - Conf. DLGP2022. - Pp. 028. - DOI: <https://doi.org/10.22323/1.429.0028>.

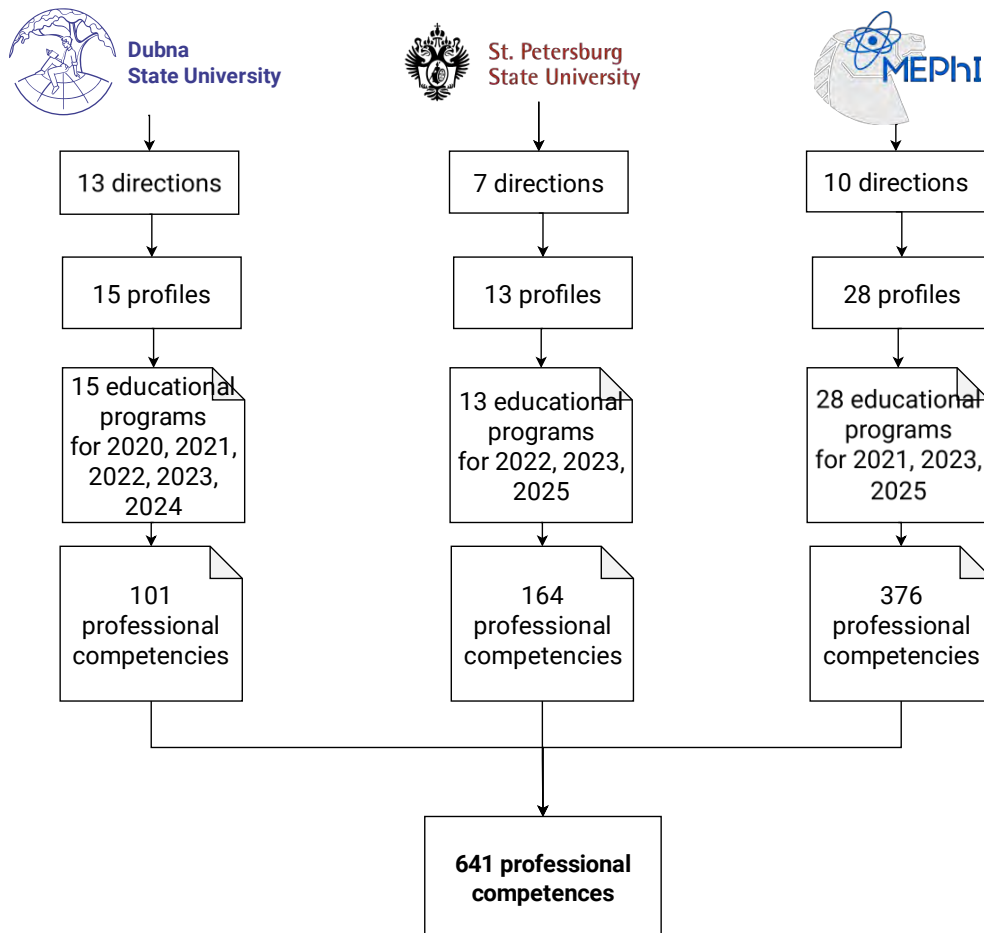
⁴ Belov, S.D., Zrellov, P.V., Ilyina, A.V., Korenkov, V.V. and Tarabrin, V.A. 2023. Using neural network language models to study the demand for higher education professional competencies on the labor market. System analysis in science and education. 3 (Nov. 2023), 13-25.

⁵ Belov, S., Ilyina, A., Korenkov, V. et al. Exploring the Relevance of Educational Skillset in the Labor Market through Natural Language Processing Techniques. Phys. Part. Nuclei 55, 584–587 (2024). <https://doi.org/10.1134/S106377962403016X>.

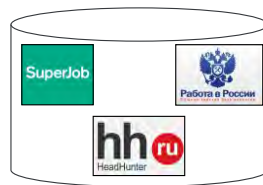
The structure of professional competence formation in the educational program



Input data. **Education**

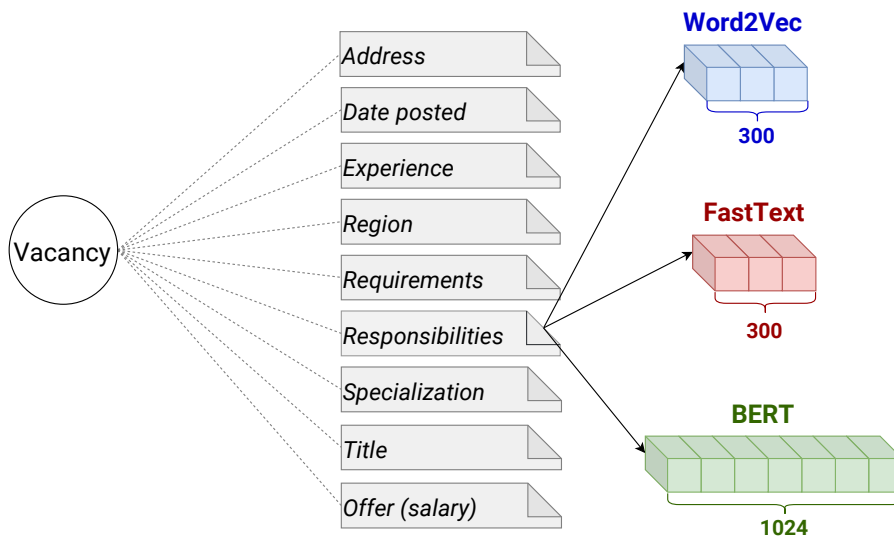


Input data. Labor market

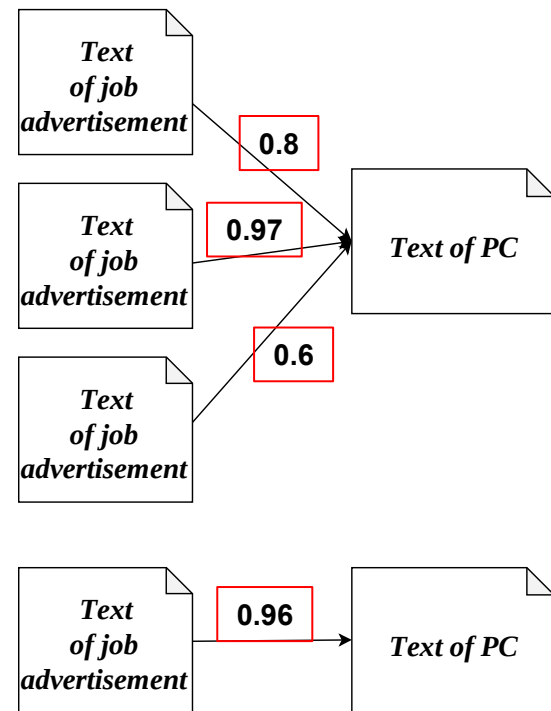


technical*
55 827
vacancies
for 2026

* technical - physics, IT & science

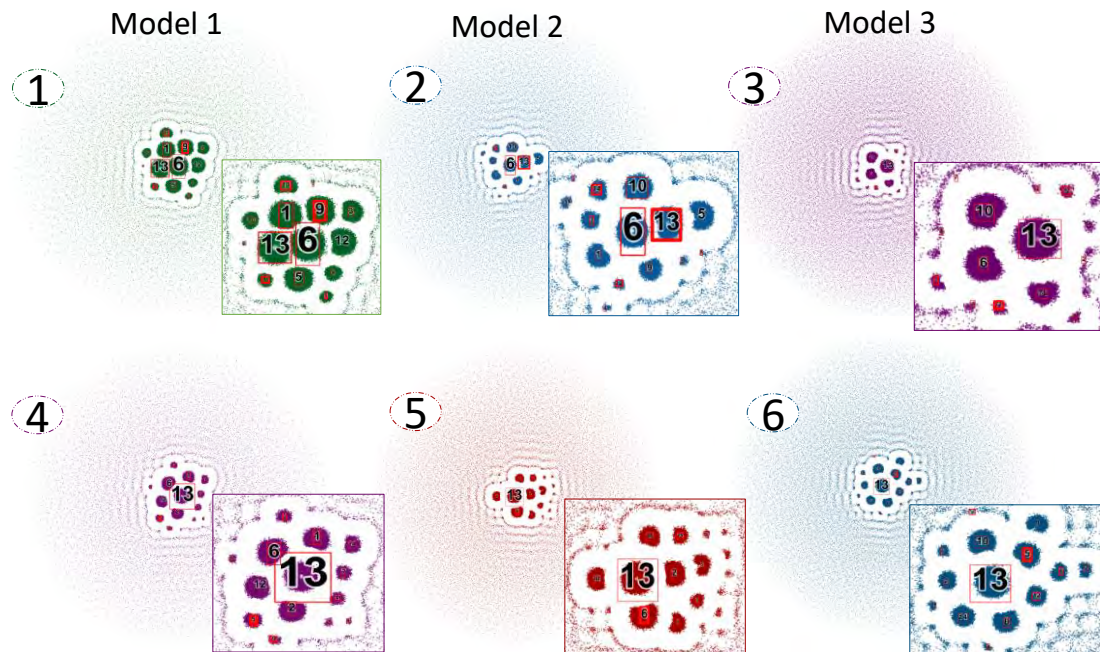


Previous results. “Vacancy – Competence” graph



Semantic similarity
between PC titles
and the "
Responsibilities"
field

Semantic similarity
between the PC
achievement
indicators and the
"Responsibilities"
field

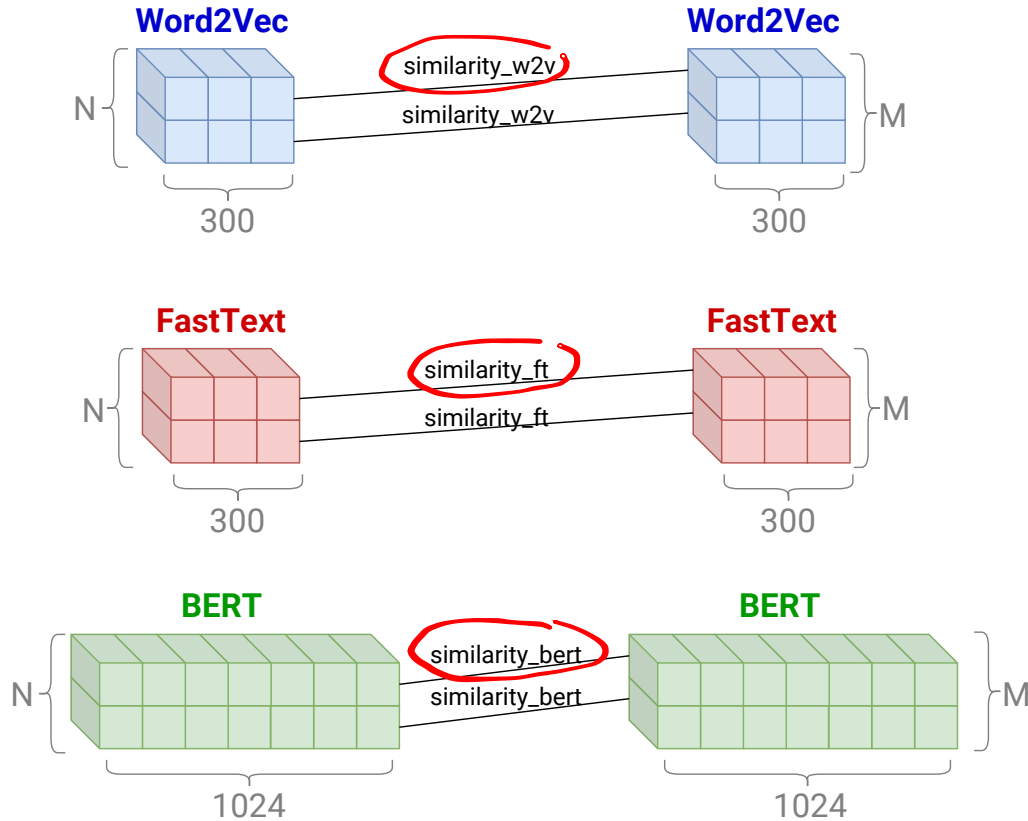


- M1**
- **Title:** [ruscorpora_1_300_10](#)
 - **Description:** Word2vec Continuous Skipgram vectors trained on full Russian National Corpus (about 250M words). The model contains 185K words.
 - **Related papers:** <https://www.academia.edu/24306935>
 - **Preprocessing:** The corpus was lemmatized and tagged with Universal PoS.
 - **Parameters:** vector size 300, window size 10

- M2**
- **Title:** [araneum_none_fasttextskipgram_300_5_2018](#)
 - **Description:** FastText Skipgram vectors trained on Russian Word Corpus (about 10⁷ words). The model contains 10B words.
 - **Related papers:** <https://arxiv.org/pdf/1801.06407.pdf>
 - **Preprocessing:** The corpus was lemmatized.
 - **Parameters:** vector size 300, window size 5

- ### M3
- **Title:** [sbert_large_mt_nlu_ru](#)
 - **Description:** BERT large model multitask (cased) for Sentence Embeddings in Russian language used to solve the problems of recognizing intent, named entity extraction, sentiment analysis, analysis of toxicity and search for similar queries.
 - **Related papers:** [https://habr.com/ru/companies/sberdevices/articles/560748/](#)
 - **Parameters:** vector size 1024, 427M parameters

The problem



*How to aggregate
to get
one value?*

N - count of professional competencies
M - count of vacancies

Vacancies & competencies

55 827 × 641 = **35 785 107 pairs**

similarities $\in (0; 1)$

we need to enter a
threshold value!

an empirically selected value: 0.8

Aggregation methods

1. Average similarity

$$\frac{sim_{w2v}(v_i, c_j) + sim_{ft}(v_i, c_j) + sim_{bert}(v_i, c_j)}{3}$$

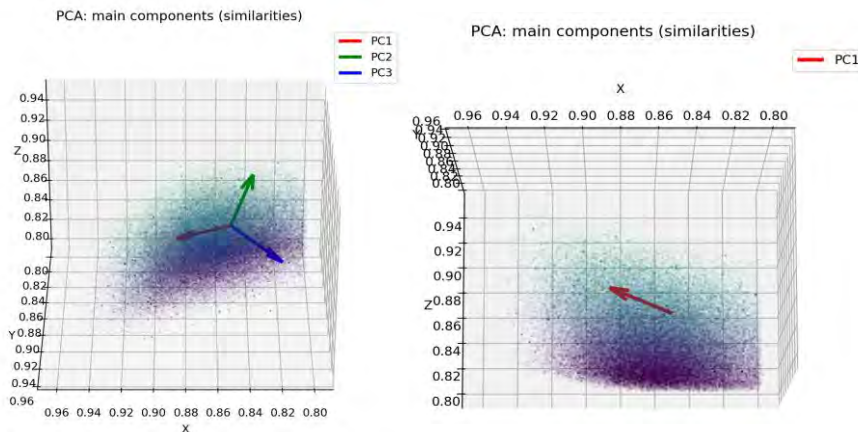
2. Similarities

aggregation using
PCA

$$\begin{matrix} sim_{w2v}(v_i, c_j), \\ sim_{ft}(v_i, c_j), \\ sim_{bert}(v_i, c_j) \end{matrix} \xrightarrow{(n_components = 1)} \begin{matrix} \text{PCA | SVD} \\ \text{MinMaxScaler} \end{matrix} \xrightarrow{(avg_min, avg_max)}$$

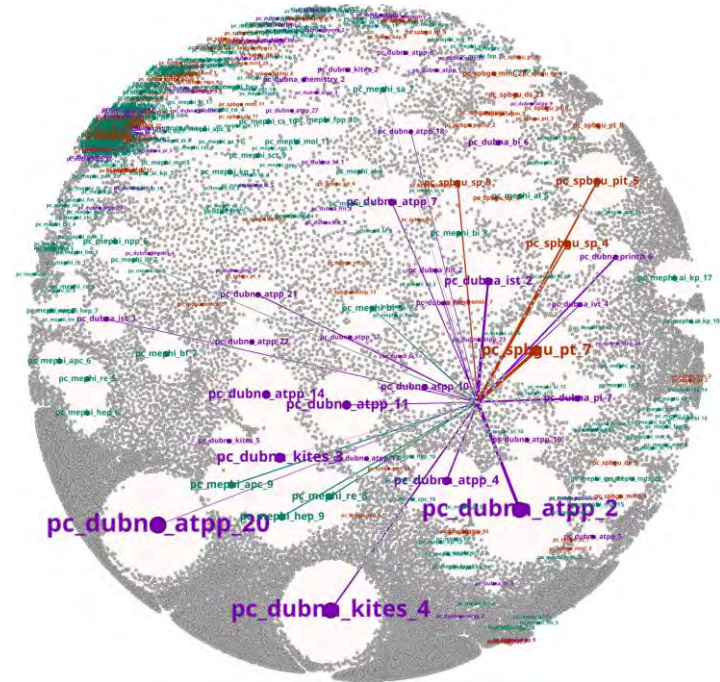
3. Similarities

aggregation using
SVD



Evaluation methods

1. Graph analysis
2. Statistical summary
3. Coefficient of variation^{5,6}
4. Correlation matrices^{7,8}
5. Kendall τ (tau) coefficient^{7,9}



⁵ Vasileios C. Pezoulas and Dimitrios I. Zaridis and Eugenia Mylona and Christos Androutsos and Kosmas Apostolidis and Nikolaos S. Tachos and Dimitrios I. Fotiadis: **Synthetic data generation methods in healthcare: A review on open-source tools and methods** // Computational and Structural Biotechnology Journal, vol.23, 2024.

⁶ Elbayumi, Mohamed et al., **Healthy-to-patients deep learning for automated left atrial segmentation and function analysis from short-axis cine CMR**, Journal of Cardiovascular Magnetic Resonance, Volume 27, DOI:10.1016/j.jocmr.2024.101495.

⁷ Shahin Atakishiyev, **Evaluation of High-Dimensional Word Embeddings using Cluster and Semantic Similarity Analysis**, M.S. thesis, Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada, 2018.

⁸ François Torregrossa, Vincent Claveau, Nihel Kooli, Guillaume Gravier, Robin Allesiardo: **On the Correlation of Word Embedding Evaluation Metrics**. LREC 2020 - 12th Conference on Language Resources and Evaluation, May 2020, Marseille, France. pp.4789 - 4797. (hal-02919006).

⁹ Mingze Dong, David Su, Harriet Kluger, Rong Fan, Yuval Kluger, **SIMV1 reveals intrinsic and spatial-induced states in spatial omics data**, bioRxiv 2023.08.28.554970; doi: <https://doi.org/10.1101/2023.08.28.554970>.

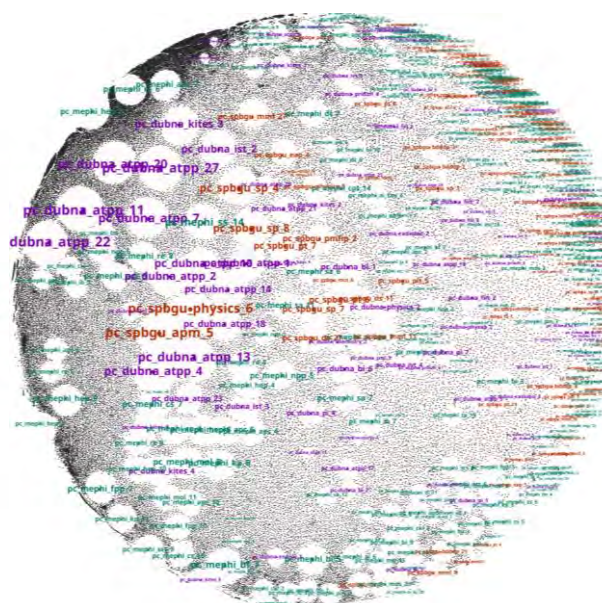
threshold=0.8

Word2Vec sim



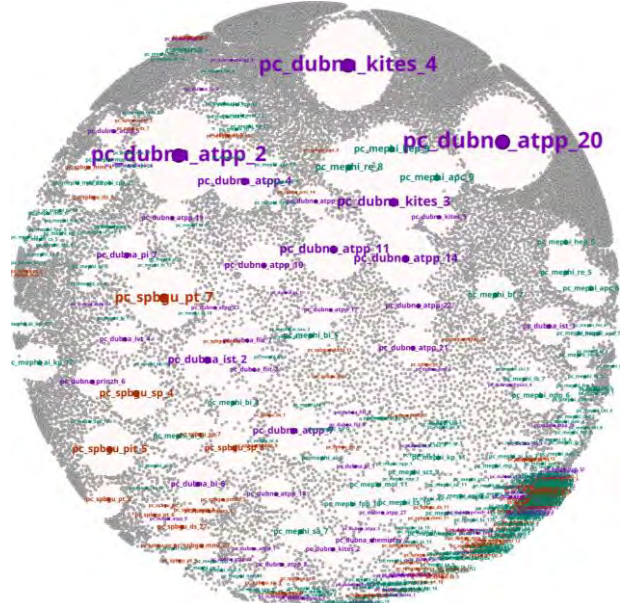
Nodes number	53 968	TOP-5	pc_dubna_attp_11	47120
			pc_dubna_attp_20	46960
			pc_dubna_attp_22	46937
Edges number	8 111 218		pc_dubna_attp_27	46821
			pc_dubna_ist_2	44997
Average degree	301	BOTT OM-5	pc_dubna_prinzh_2	2
			pc_mephi_ai_kp_6	2
			pc_spbg_u_mml_31	1
			pc_spbg_u_bdddp_20	1
			pc_spbg_u_ds_31	1

FastText sim



Nodes number	50 654	TOP-5	pc_dubna_atpp_22	36035
			pc_dubna_atpp_11	35667
			pc_spbg_u_apm_5	34712
Edges number	2 565 988		pc_dubna_atpp_27	34046
			pc_spbg_u_physics_6	33799
Average degree	101	BOTT OM-5	pc_mephi_bf_4	1
			pc_spbg_u_ds_19	1
			pc_mephi_ai_kp_10	1
			pc_mephi_mds_2	1
			pc_mephi_pm_2	1

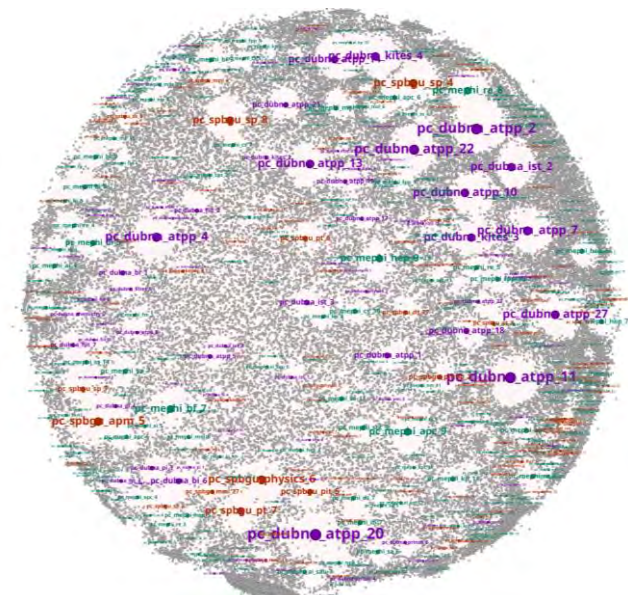
BERT sim



Nodes number	23 058	TOP-5	pc_dubna_atpp_20	11266
			pc_dubna_atpp_2	10370
			pc_dubna_kites_4	10360
Edges number	185 450		pc_spgbu_pt_7	6017
			pc_dubna_kites_3	5576
Average degree	16	BOTT OM-5	pc_mephi_kp_10	1
			pc_mephi_fm_6	1
			pc_mephi_tp_8	1
			pc_dubna_energy_1	1
			pc_mephi_mp_14	1

threshold=0.8

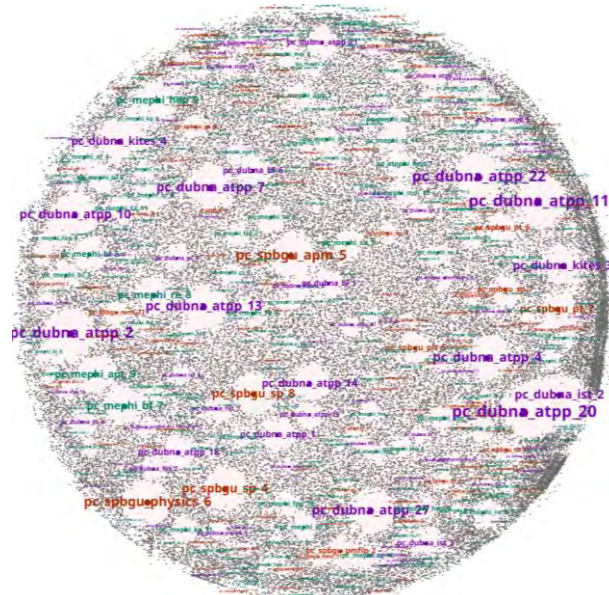
AVG sim



Nodes number	44 976
Edges number	1 261 092
Average degree	56

TOP-5	pc_dubna_atpp_20	29016
	pc_dubna_atpp_11	27391
	pc_dubna_atpp_2	26542
	pc_dubna_atpp_22	24592
	pc_dubna_atpp_4	21942
BOTT	pc_mephi_cs_12	1
OM-5	pc_dubna_pi_5	1
	pc_mephi_fpp_14	1
	pc_mephi_tp_10	1
	pc_spbgu_ds_20	1

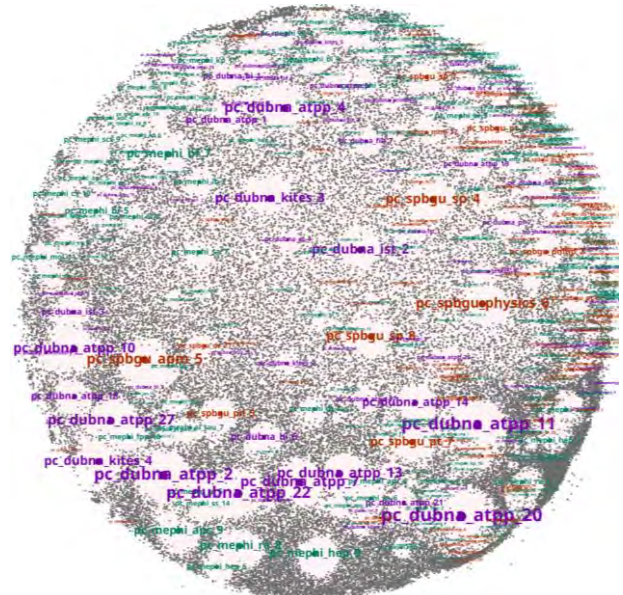
PCA sim



Nodes number	44 866
Edges number	1 233 719
Average degree	55

TOP-5	pc_dubna_atpp_20	28055
	pc_dubna_atpp_11	27056
	pc_dubna_atpp_22	25379
	pc_dubna_atpp_2	24992
	pc_spbgu_apm_5	22466
BOTT	pc_mephi_mds_13	1
OM-5	pc_mephi_fpp_14	1
	pc_mephi_tp_10	1
	pc_spbgu_ds_21	1
	pc_mephi_mol_15	1

SVD sim



Nodes number	44 241
Edges number	1 162 468
Average degree	53

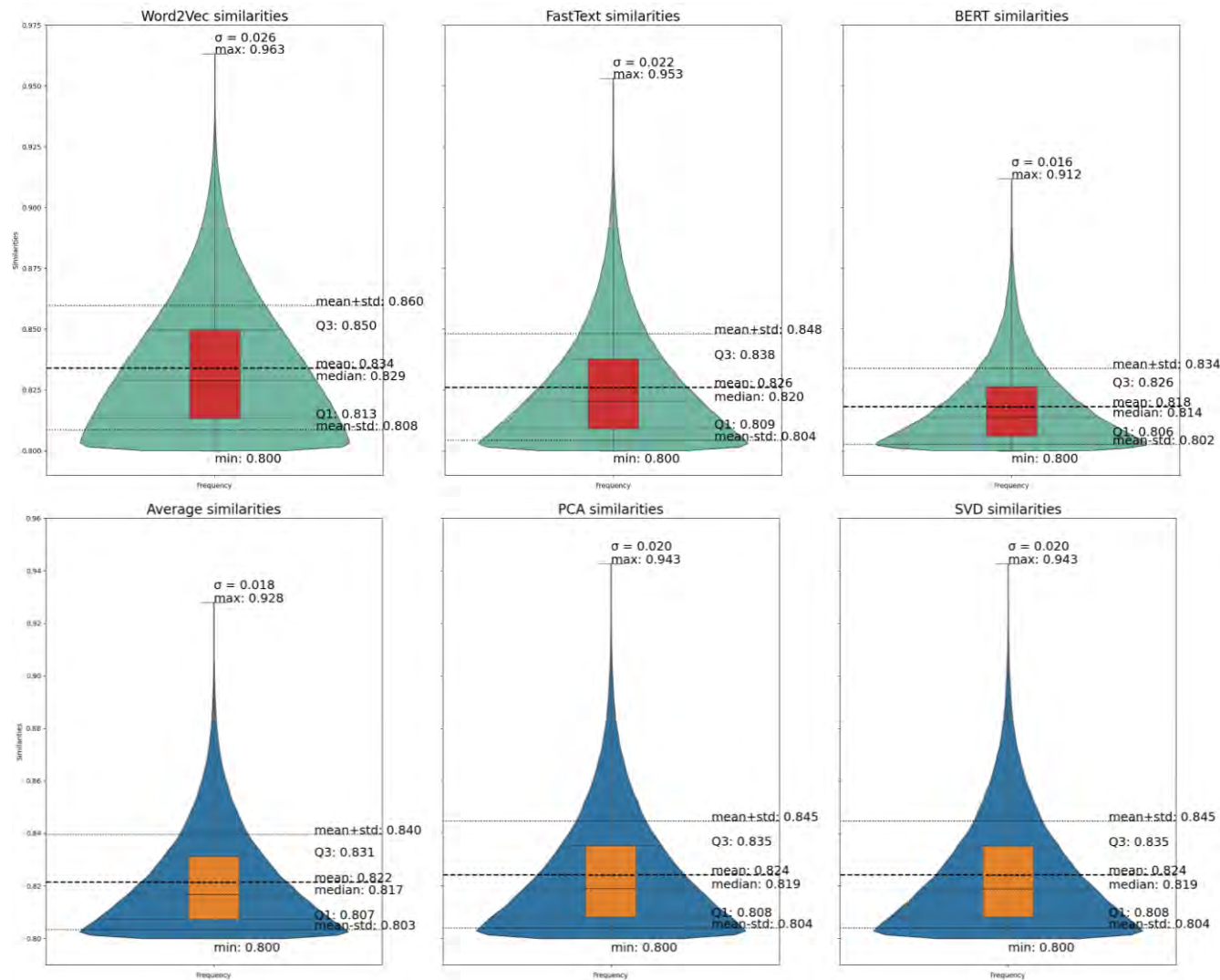
TOP-5	pc_dubna_atpp_20	27705
	pc_dubna_atpp_11	26266
	pc_dubna_atpp_2	24895
	pc_dubna_atpp_22	24106
	pc_dubna_atpp_4	20826
BOTT	pc_mephi_mol_13	1
OM-5	pc_mephi_kp_13	1
	pc_mephi_cps_24	1
	pc_mephi_cs_12	1
	pc_spbgu_pmi_11	1

Evaluation methods

1. Graph analysis
2. Statistical summary
3. Coefficient of variation
4. Correlation matrices
5. Kendall τ (tau) coefficient

The cosine similarity distributions obtained by the studied models are characterized by positive asymmetry:

- the average value exceeds the median, which indicates the presence of pairs with high similarity values.
- The largest spread of values is observed for the Word2Vec model ($\sigma = 0.026$), while the BERT model demonstrates the most compact distribution ($\sigma = 0.016$).**
- The interquartile range is also maximum for Word2Vec and minimum for BERT, indicating that **Word2Vec generates more variable similarity scores and BERT generates more concentrated scores.**
- The width of the violin plots reflects the distribution density of the values and shows that the majority of observations are concentrated in the range of 0.80-0.84 for all models.



All aggregation methods exhibit positive distribution asymmetry:

- The means exceed the medians.
- Simple averaging produces the most compact distribution ($\sigma = 0.018$), smoothing out the differences between objects.**
- PCA and SVD methods demonstrate almost identical distributions and allow us to obtain more contrasting similarity estimates: the standard deviation increases to 0.020, the interquartile range increases from 0.024 to 0.027, and the maximum values reach 0.943, compared to 0.928 in the case of simple averaging.
- PCA and SVD better preserve the variability of the original model estimates and provide a more pronounced separation of the "competency-vacancy" pairs based on their degree of similarity.**

Evaluation methods

1. Graph analysis
2. Statistical summary
3. Coefficient of variation
4. Correlation matrices
5. Kendall τ (tau) coefficient

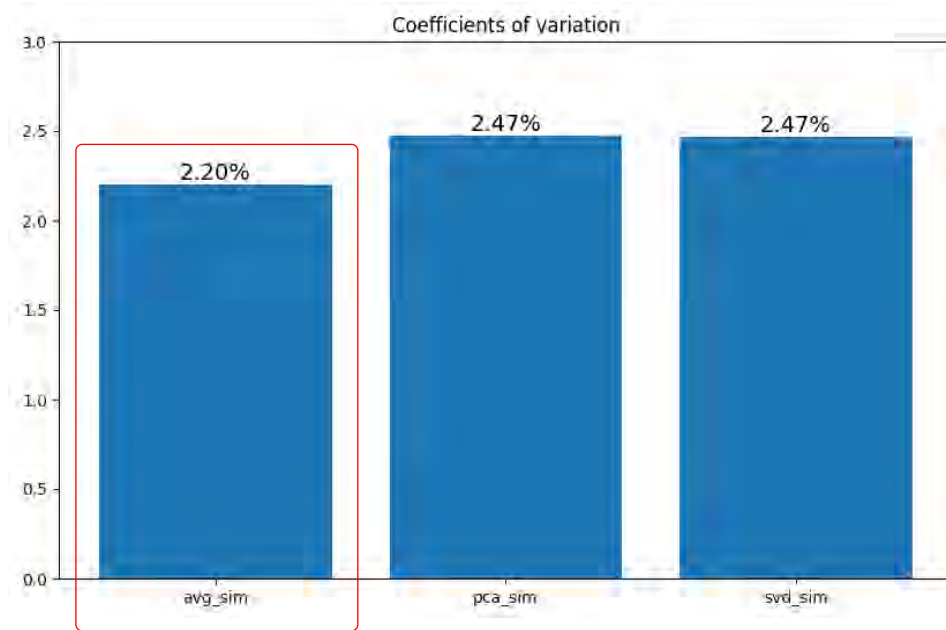
The higher the coefficient of variation, the greater the spread of the data relative to the mean.

$$CV = \left(\frac{\sigma}{\mu} \right) * 100\%$$

A low coefficient value indicates a more homogeneous data set.

All coefficients of variation are close to zero, which indicates sufficient homogeneity of the obtained proximity values.

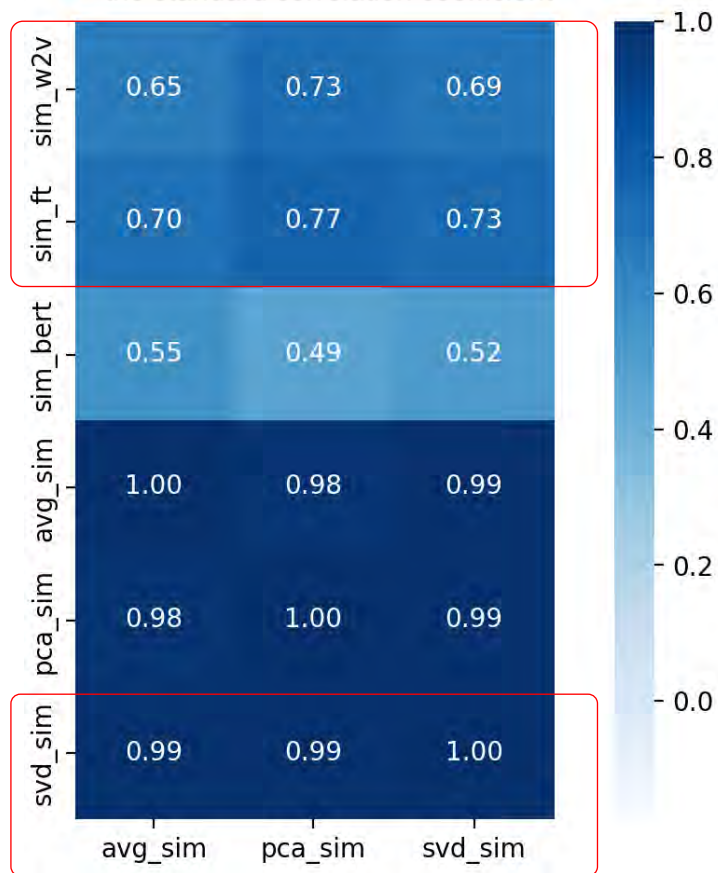
The lowest coefficient of variation is for the **averaging method (Average similarity) (2.20%)**, i.e. the aggregated results obtained by this method are **the most homogeneous**.



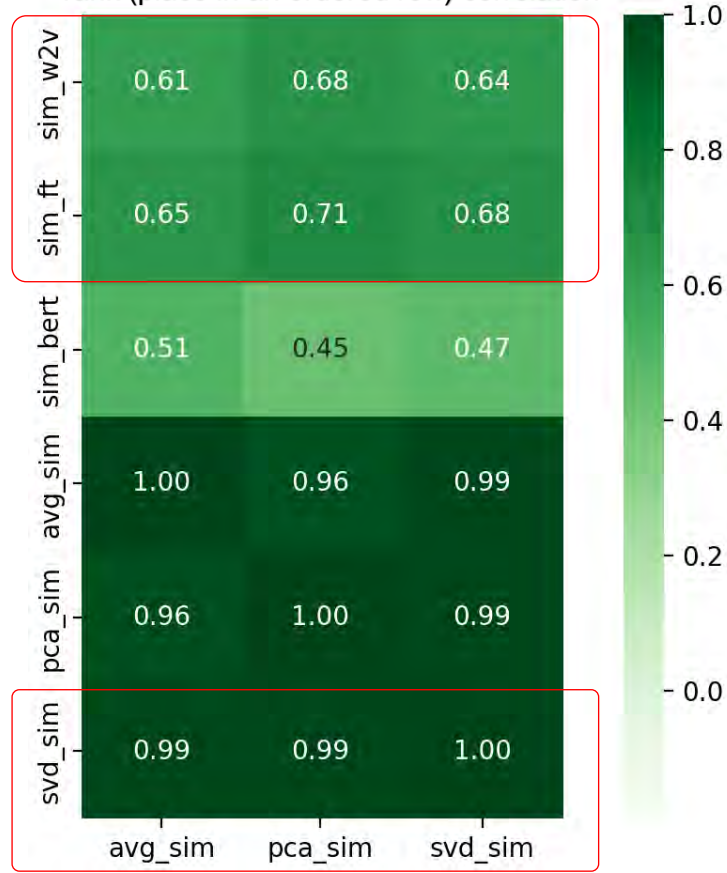
Evaluation methods

1. Graph analysis
2. Statistical summary
3. Coefficient of variation
4. Correlation matrices
5. Kendall τ (tau) coefficient

Pearson correlation matrix:
the standard correlation coefficient



Spearman correlation matrix:
rank (place in an ordered row) correlation



Evaluation methods

1. Graph analysis
2. Statistical summary
3. Coefficient of variation
4. Correlation matrices
5. Kendall τ (tau) coefficient

The **Kendall τ (tau) coefficient** measures the consistency of ranks from two experts.

The value ranges from 0 (complete disagreement) to 1 (perfect agreement).

- **τ (tau)** – correlation between **two sets of ranks**.

The τ formula:

$$\tau = \frac{C - D}{0.5 * n * (n - 1)}$$

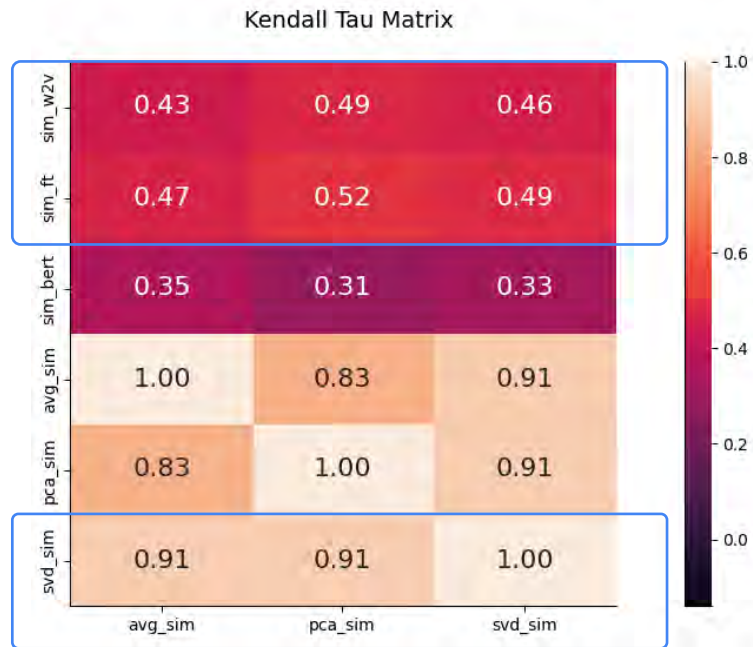
where:

- n is the total number of observations (pairs),
- C (concordant) is the number of matched pairs (the ranks of both features change in the same direction),
- D (discordant) is the number of mismatched pairs (the ranks change in opposite directions).

- All three aggregation methods produce almost the same order of the most relevant pairs.
- BERT rankings differ significantly from ensemble rankings.
- FastText is closer to ensembles.

The Kendall coefficients between the aggregation methods are in the range of 0.83-0.91, which indicates the high stability of the final ranking of the 'vacancy-competence' pairs to the choice of the method of combining similarity scores.

At the same time, BERT shows lower consistency with ensemble methods ($\tau = 0.31$ -0.35), while Word2Vec and FastText show moderate consistency ($\tau = 0.43$ -0.52). This suggests that the aggregate scores inherit the similarity structure generated by Word2Vec and FastText to a greater extent than BERT.



Results

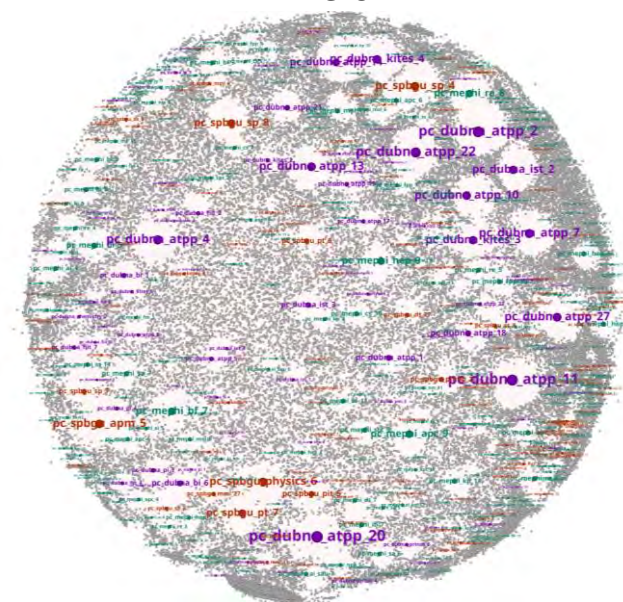
1. All considered ensemble methods demonstrate similar results.
2. The spread (standard deviation) and coefficient of variation in the simple averaging method show the smallest values, which indicates greater homogeneity of the results.
3. The method of simple averaging of similarities has the lowest computational costs compared to other considered methods.

Results. Portrait of a graduate's demand

Count of edges (TOP-3)

University	↓ By directions	↓ By profiles	↓ By regions	↓ By offer (RUB)
National Research Nuclear University MEPhI	Nuclear physics and technologies 156 440 edges	Nuclear physics and cosmophysics 56 392 edges	Moscow 127 986 edges	100 000 33 088 edges
	Applied mathematics and physics 122 237 edges	Elementary particle physics and cosmology 50 870 edges	Saint-Petersburg 47 045 edges	150 000 32 879 edges
	Business informatics 67 225 edges	Business Informatics in the digital economy 49 178 edges	Moscow region 32 199 edges	120 000 24 532 edges
St. Petersburg State University	Applied mathematics and informatics 73 579 edges	Modern programming 58 128 edges	Moscow 73 963 edges	150 000 16 427 edges
	Mathematics and computer science 47 645 edges	Programming technologies 39 344 edges	Saint-Petersburg 24 024 edges	100 000 15 013 edges
	Mathematical support and administration of information systems 39 344 edges	Physics 21 794 edges	Moscow region 13 196 edges	200 000 11 489 edges
Dubna State University	Automation of technological processes and productions 303 889 edges	Intelligent control systems for robotic technological processes 303 889 edges	Moscow 108 814 edges	100 000 32 405 edges
	Design and technology of electronic devices 54 885 edges	Design and technology of intelligent radio-electronic devices 54 885 edges	Saint-Petersburg 43 420 edges	150 000 30 007 edges
	Information systems and technologies 34 103 edges	Information system security 34 103 edges	Moscow region 33 866 edges	120 000 23 863 edges

AVG sim



Conclusion

The final aggregation strategy for combining similarity scores from the selected models can be any of the three methods considered, as they show similar results.

However, if it is necessary to achieve the accuracy and justification of the chosen method, it is worth stopping at the method of **simple average similarity** – thus, it is the closest approach to their “consensus judgment”.

The results obtained coincide with the results obtained when working with other datasets of vacancies and competencies (reported at the GRID-2025 conference).

Thank you for your attention!

Reserve slides

```
pd.DataFrame(df.iloc[0, :])
```

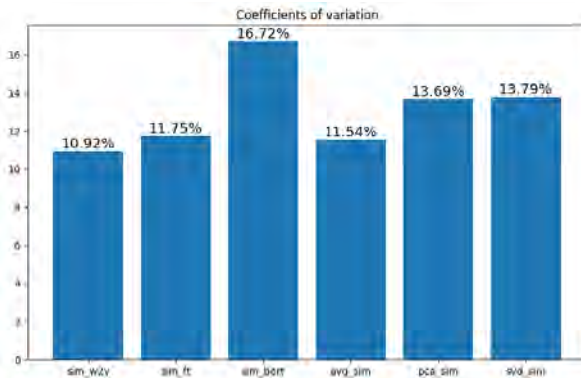
0

address	Нижний Новгород, Московское шоссе, 31А
date_posted	2022-01-17T14:40:57+0300
experience	От 1 года до 3 лет
platform	hh
profArea	Информационные технологии, интернет, телеком
region	Нижегородская область
requirements	Опыт от 1 года. Полная занятость, полный день (З/П по результатам собеседования). ===== Ускорь карьерный и профессиональный рост! У нас много современных и перспективных бизнес-направлений - работа с цифровыми системами (ЕГАИС и др.), автоматизация торговли, защита информации, альтернативная энергетика (солнечные электростанции и не только). &quot;ЦЭК&quot; стабильно...
responsibilities	: - Создать сайт с нуля на популярном движке, &quot;натянуть дизайн&quot;, красиво и адаптивно сверстать его; - Внести правки в чужой код и не сломать сайт; - Сверстать лендинг с анимацией, стильное емейл-письмо; - хорошо знаешь HTML, PHP, CSS, Js и т.д. (необходимый веб-инструментарий). Нам важен твой скилл, а не гарвардская степень и стаж в мега-корпорациях. Звони прямо сейчас! Мы пригласим ...
schedule	Полный день
specialization	Web мастер
title	web-программист html-верстальщик
conditions	None

ID of PC	Name	Direction	Profile	University	Year
pc_dubna_atpp_20	Способность выполнять работы по регламентному техническому, эксплуатационному обслуживанию оборудования, средств и систем автоматизации и управления, а также средств программного обеспечения	Автоматизация технологических процессов и производств	Интеллектуальные системы управления роботизированными технологическими процессами	Государственный университет «Дубна»	2021
pc_dubna_atpp_11	Способность участвовать в разработке проектной и рабочей технической документации в области автоматизации технологических процессов и их производств, эксплуатационному обслуживанию, управлению жизненным циклом и ее продукции качеством	Автоматизация технологических процессов и производств	Интеллектуальные системы управления роботизированными технологическими процессами	Государственный университет «Дубна»	2021
pc_dubna_atpp_2	Способность участвовать в работах по моделированию продукции, технологических процессов, производств, средств и систем автоматизации, контроля, диагностики, испытаний и управления процессами, жизненным циклом продукции и ее качеством с использованием современных средств автоматизированного проектирования, по разработке алгоритмического и программного обеспечения средств и систем автоматизации и управления процессами	Автоматизация технологических процессов и производств	Интеллектуальные системы управления роботизированными технологическими процессами	Государственный университет «Дубна»	2021
pc_mephi_bf_7	Способен анализировать исходные данные проектирования, участвовать в разработке, подготовке и оформлении проектной документации с учетом норм радиационной и экологической безопасности	Физика	Биофизика	Национальный исследовательский ядерный университет "МИФИ"	2021
pc_mephi_hep_9, pc_mephi_re_8	Способен к выполнению работ по стандартизации и подготовке к сертификации технических средств, систем, оборудования и материалов	Ядерные физика и технологии	Физика элементарных частиц и космология, Радиационная экология и обеспечение безопасности человека и окружающей среды	Национальный исследовательский ядерный университет "МИФИ"	2021
pc_spbgu_apm_5, pc_spbgu_physics_6	Способен понимать принципы составления проектов работ в избранной области, представлять экономические аспекты проектной деятельности, участвовать в сертификации технических средств (оборудования, алгоритмов, программных продуктов) и/или в подготовке материалов для защиты объектов интеллектуальной собственности	Прикладная математика и физика, Физика	Прикладные математика и физика, Физика	Санкт-Петербургский государственный университет	2022, 2023, 2024, 2025
pc_spbgu_sp_4	Способен профессионально решать задачи производственной и технологической деятельности с учетом современных достижений науки и техники, включая: разработку алгоритмических и программных решений в области системного и прикладного программирования; разработку математических, информационных и имитационных моделей по тематике выполняемых исследований; создание информационных ресурсов глобальных сетей, прикладных баз данных; разработку тестов и средств тестирования систем и средств на соответствие стандартам и исходным требованиям; разработку эргономичных человеко-машинных интерфейсов; разработку систем среднего и крупного масштаба сложности	Прикладная математика и информатика	Современное программирование	Санкт-Петербургский государственный университет	2022, 2023, 2024

ID of PC	Name	Direction	Profile	University	Year
pc_dubna_atpp_20	The ability to carry out maintenance work on equipment, automation and control systems, as well as software tools.	Automation of technological processes and productions	Intelligent control systems for robotic technological processes	Dubna State University	2021
pc_dubna_atpp_11	Ability to participate in the development of project and working technical documentation in the field of automation of technological processes and their production, operational maintenance, lifecycle management, and product quality	Automation of technological processes and productions	Intelligent control systems for robotic technological processes	Dubna State University	2021
pc_dubna_atpp_2	The ability to participate in work related to the modeling of products, technological processes, production systems, tools, and systems for automation, control, diagnostics, testing, and process management, including the product life cycle and its quality, using modern computer-aided design tools, as well as in the development of algorithmic and software solutions for automation and process management systems	Automation of technological processes and productions	Intelligent control systems for robotic technological processes	Dubna State University	2021
pc_mephi_bf_7	Capable of analyzing initial design data, participating in the development, preparation, and documentation of project materials taking into account radiation and environmental safety standards	Physics	Biophysics	National Research Nuclear University MEPhI	2021
pc_mephi_hep_9, pc_mephi_re_8	Capable of performing work on the standardization and preparation for certification of technical means, systems, equipment, and materials	Nuclear physics and technologies	Particle physics and cosmology, Radiation ecology and ensuring the safety of humans and the environment	National Research Nuclear University MEPhI	2021
pc_spbgu_apm_5, pc_spbgu_physics_6	Capable of understanding the principles of developing work projects in the chosen field, representing the economic aspects of project activities, participating in the certification of technical means (equipment, algorithms, software products) and/or in the preparation of materials for the protection of intellectual property objects	Applied mathematics and physics, Physics	Applied mathematics and physics, Physics	St. Petersburg State University	2022, 2023, 2024, 2025
pc_spbgu_sp_4	Capable of professionally addressing tasks of production and technological activities taking into account modern achievements in science and technology, including: the development of algorithmic and software solutions in the field of system and application programming; the development of mathematical, informational, and simulation models related to the topics of conducted research; the creation of information resources for global networks, application databases; the development of tests and testing tools for systems and tools to ensure compliance with standards and initial requirements; the development of ergonomic human-machine interfaces; the development of medium- and large-scale complex systems.	Applied mathematics and informatics	Modern programming	St. Petersburg State University	2022, 2023, 2024

Statistics for the full dataset (no threshold filtering)



W (concordance) – the agreement of many experts at once.

W = 1 means that all the experts arranged the objects in the same order. This indicates complete agreement.

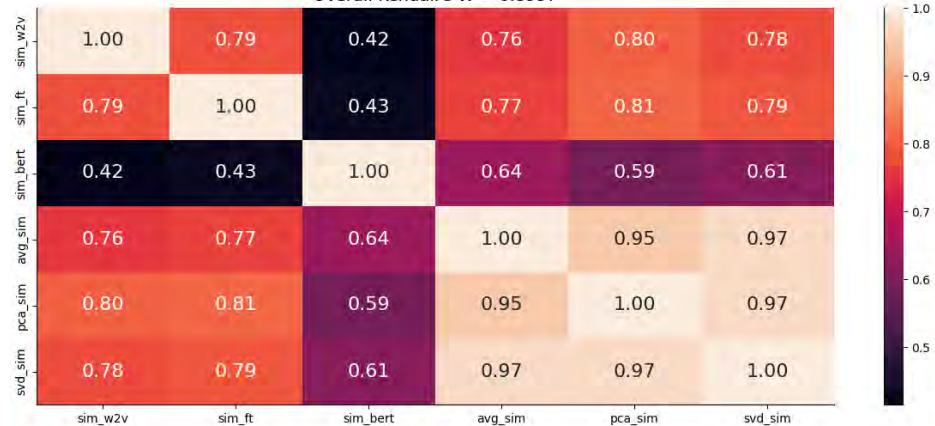
W = 0 means that the arrangements are unrelated, as if the ranks were randomly selected.

Intermediate values indicate partial agreement, where there is a general trend but the judges disagree on specific details.

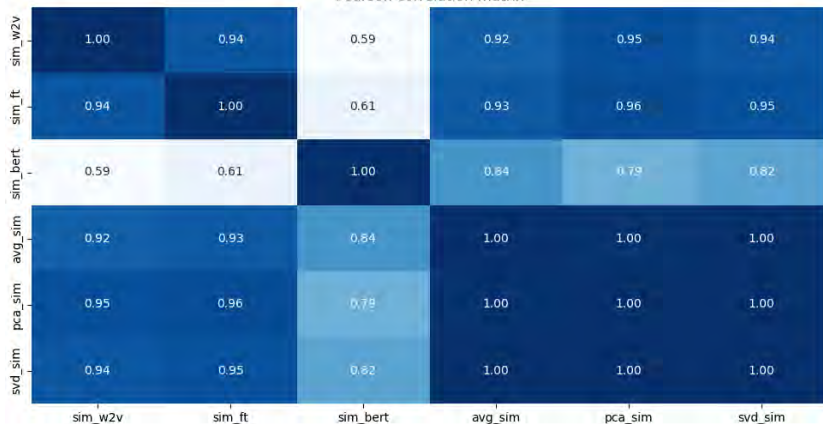
$$W = \frac{12S}{m^2(n^3 - n)}$$

where S is the sum of squared deviations of the sums of ranks from their mean, m is the number of experts, n is the number of objects.

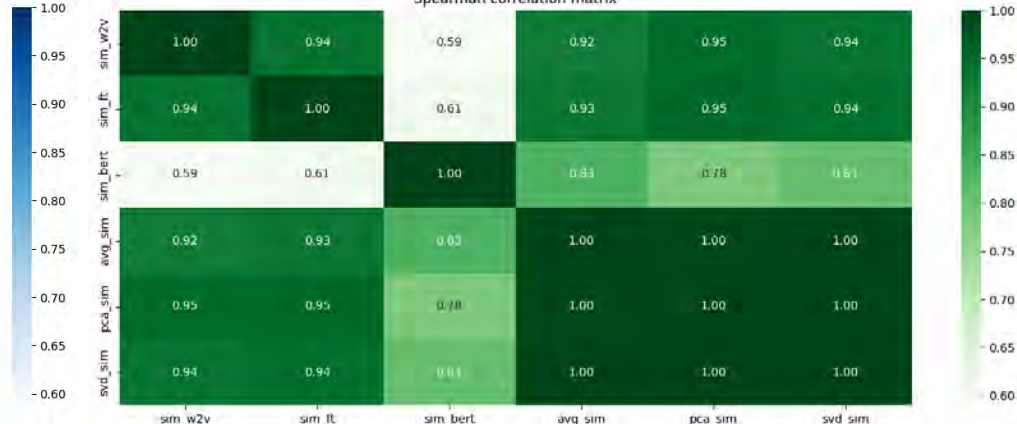
Kendall Tau Matrix
Overall Kendall's W = 0.8997

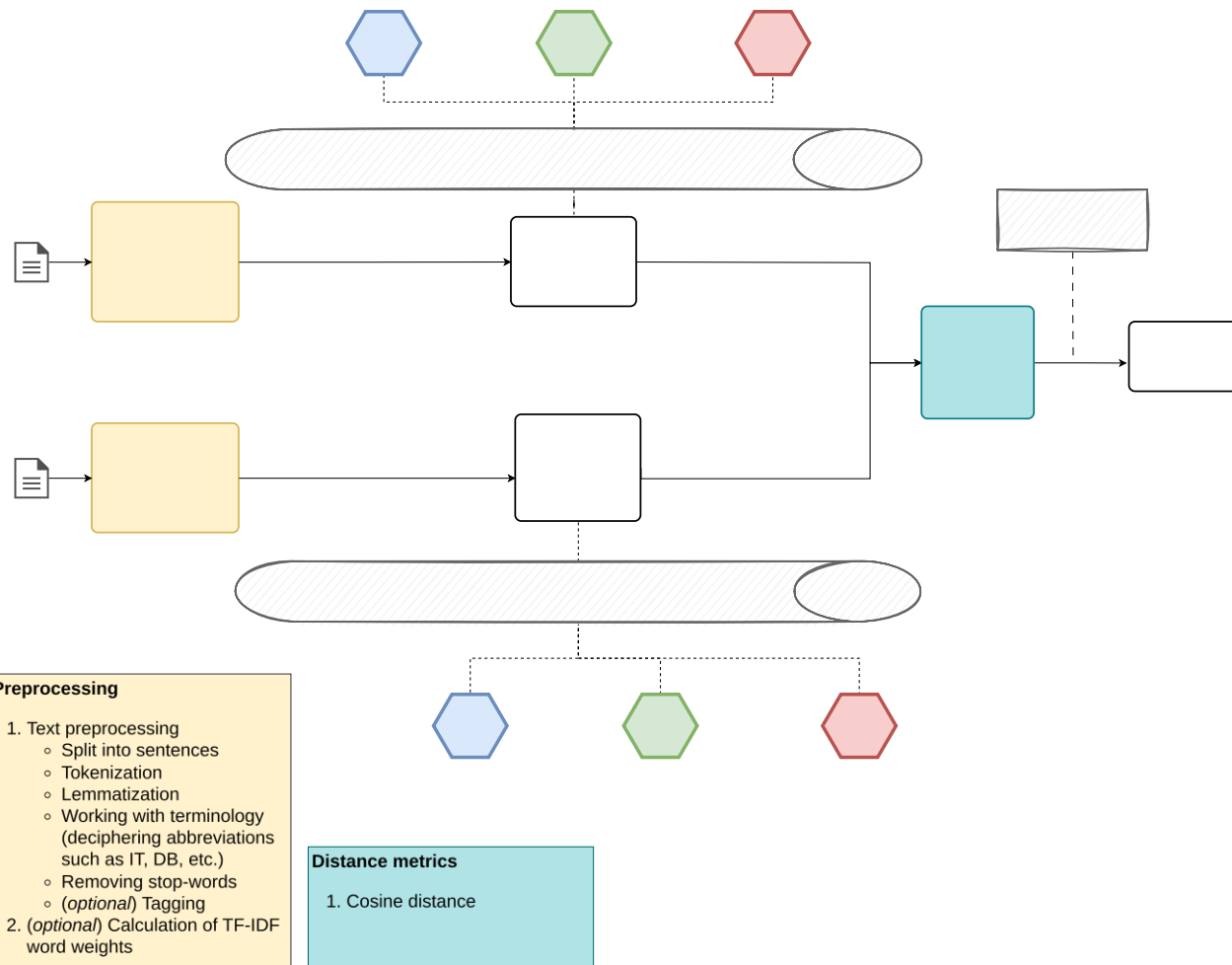


Pearson correlation matrix



Spearman correlation matrix





M1

- **Title:** ruscorpora 1 300 10
- **Description:** Word2vec Continuous Skipgram vectors trained on full Russian National Corpus (about 250M words). The model contains 185K words.
- **Related papers:** <https://www.academia.edu/24306935>
- **Preprocessing:** The corpus was lemmatized and tagged with Universal PoS.
- **Parameters:** vector size 300, window size 10

M2

- **Title:** araneum none fasttextskipgram 300 5 2018
- **Description:** FastText Skipgram vectors trained on Russian Web Corpus (about 10^7 words). The model contains 10B words.
- **Related papers:** <https://arxiv.org/pdf/1801.06407.pdf>
- **Preprocessing:** The corpus was lemmatized.
- **Parameters:** vector size 300, window size 5

M3

- **Title:** sbert large mt nlu ru
- **Description:** BERT large model multitask (cased) for Sentence Embeddings in Russian language used to solve the problems of recognizing intent, named entity extraction, sentiment analysis, analysis of toxicity and search for similar queries.
- **Related papers:** <https://habr.com/ru/companies/sberdevices/articles/560748/>
- **Parameters:** vector size 1024, 427M parameters