JOINT INSTITUTE
FOR NUCLEAR RESEARCH
ОБЪЕДИНЕННЫЙ ИНСТИТУТ
ЯДЕРНЫХ ИССЛЕДОВАНИЙ

JINR MLIT

Сравнительный анализ и оценка согласованности методов интеграции семантических представлений текстов

Comparative analysis and evaluation of the consistency of methods for integrating semantic representations of texts

Anna Ilina[1,a], Petr Zrelov[1,b]

[a]annailina@jinr.ru, [b]zrelov@jinr.ru

[1]*Meshcheryakov Laboratory of Information Technologies of the Joint Institute for Nuclear Research*

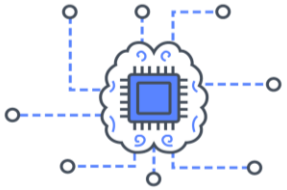
Motivation

Analyzing the link between education and labor markets is a key research priority, as technological changes and rising skill demands require adapting educational programs.

A mismatch between graduate skills and economic needs leads to labor shortages and unemployment [1-3].

Understanding these mechanisms is crucial for developing effective education policies and workforce demand forecasting tools.

Recent years have seen widespread adoption of AI tools like **machine learning, NLP, semantic methods, ontologies** [4-13]. There's growing emphasis on developing integrated decision support systems and **end-to-end data analytics platforms** [6, 14-17].

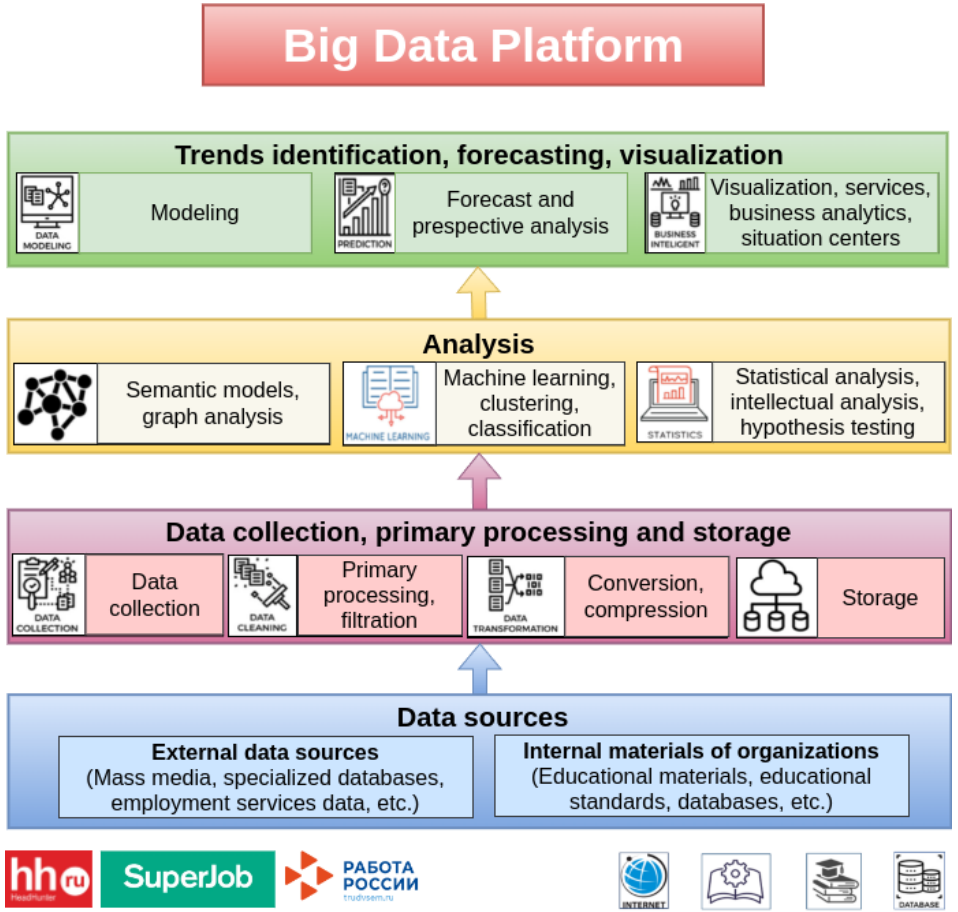

Machine Learning


Neural Networks


Natural Language Processing

Papers [4–8] present the authors' work on the creation and development of an **analytical platform based on Big Data solutions and technologies that implements a complete data processing cycle** – from collection and storage to semantic analysis and services for visualising results and making decisions in the field of monitoring and analysing the labour market and matching employers' staffing needs with the level of training of specialists in the Russian Federation.

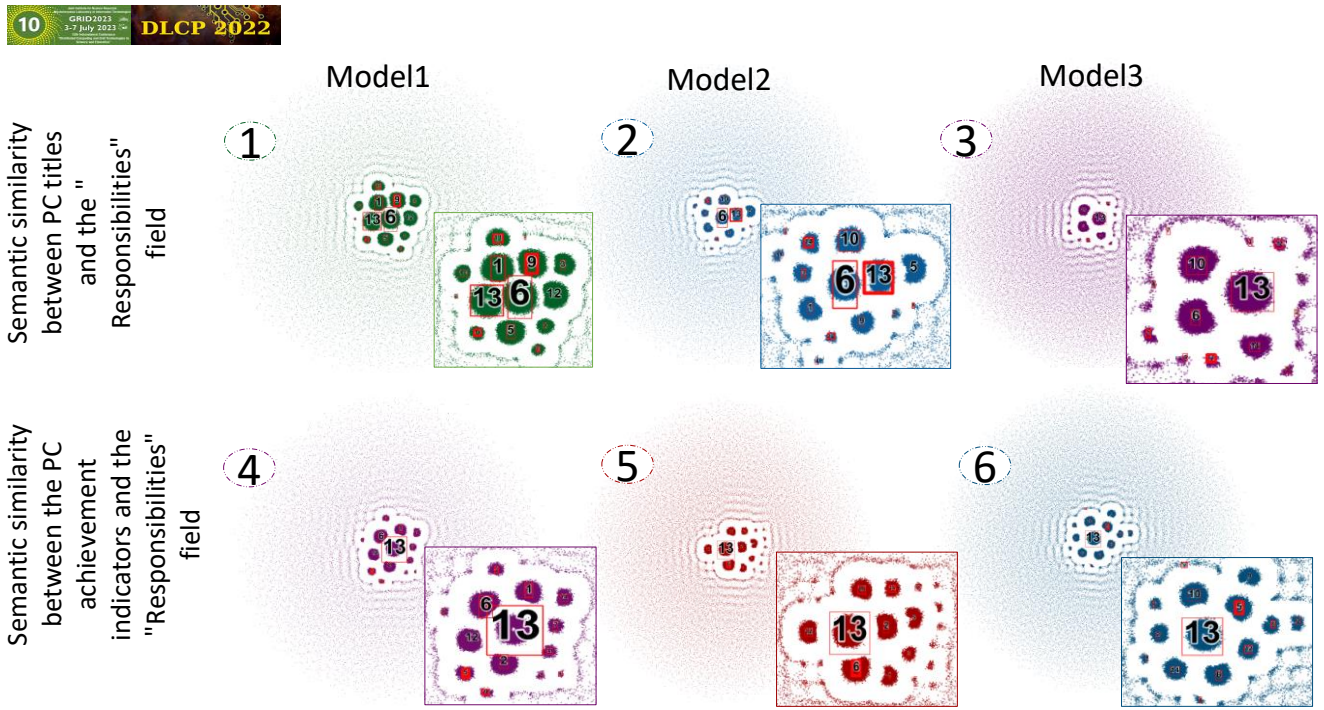
This work presents the results related to the development of the methods proposed in [4–8].

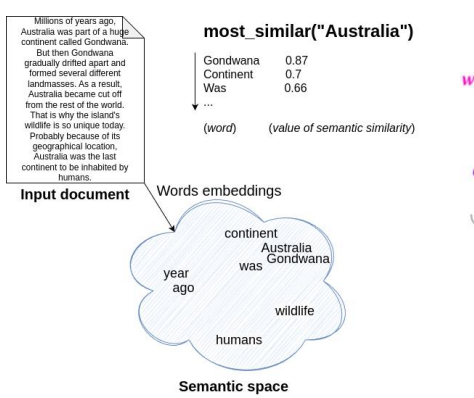


Task description

In papers [7-8], the proposed methodology was verified by comparative analysis of graphs reflecting the connections between ‘Professional Competence’ and ‘Vacancy’ using three embedding models of different architectures (Word2Vec, FastText, BERT).

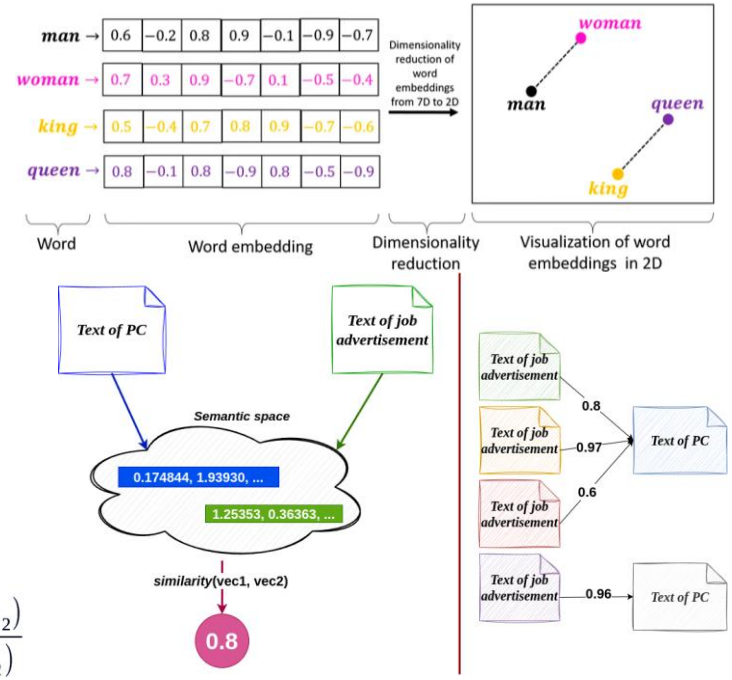
This work aims **to investigate ways of aggregating such representations and to conduct a comparative analysis of various methods for assessing their consistency and stability**. The results obtained can be used to improve the stability of models for assessing the semantic proximity of texts when using ensembles of embeddings.



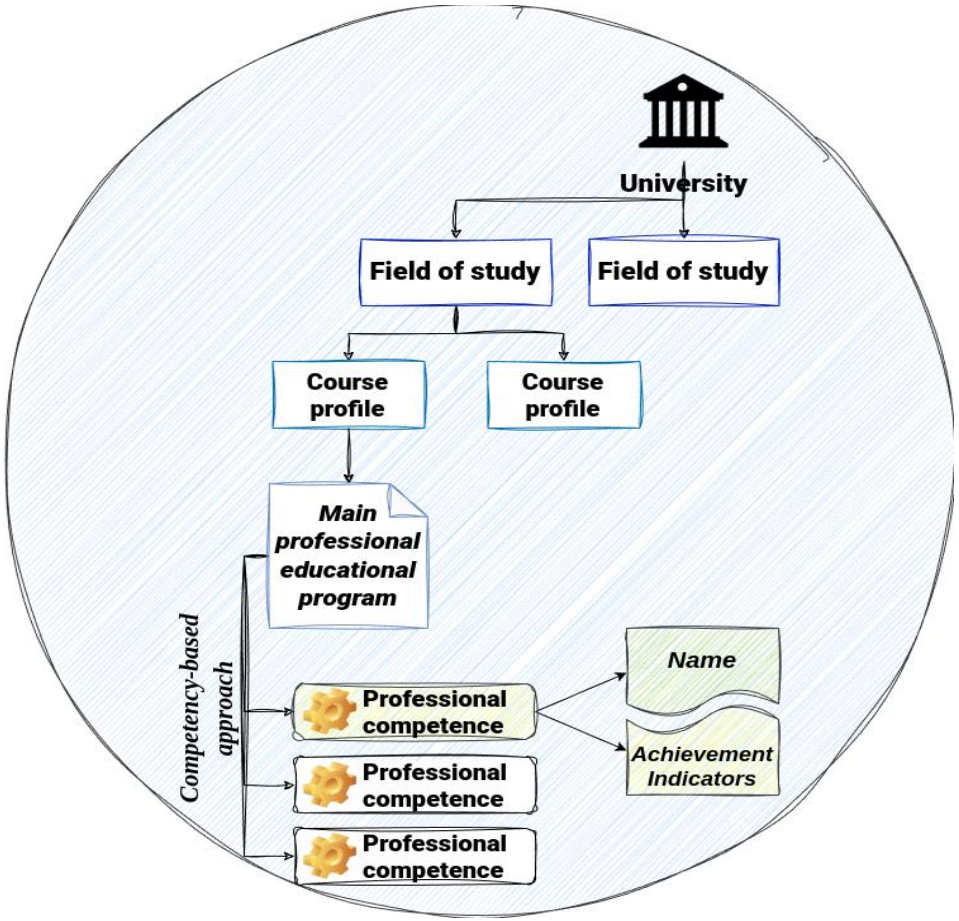


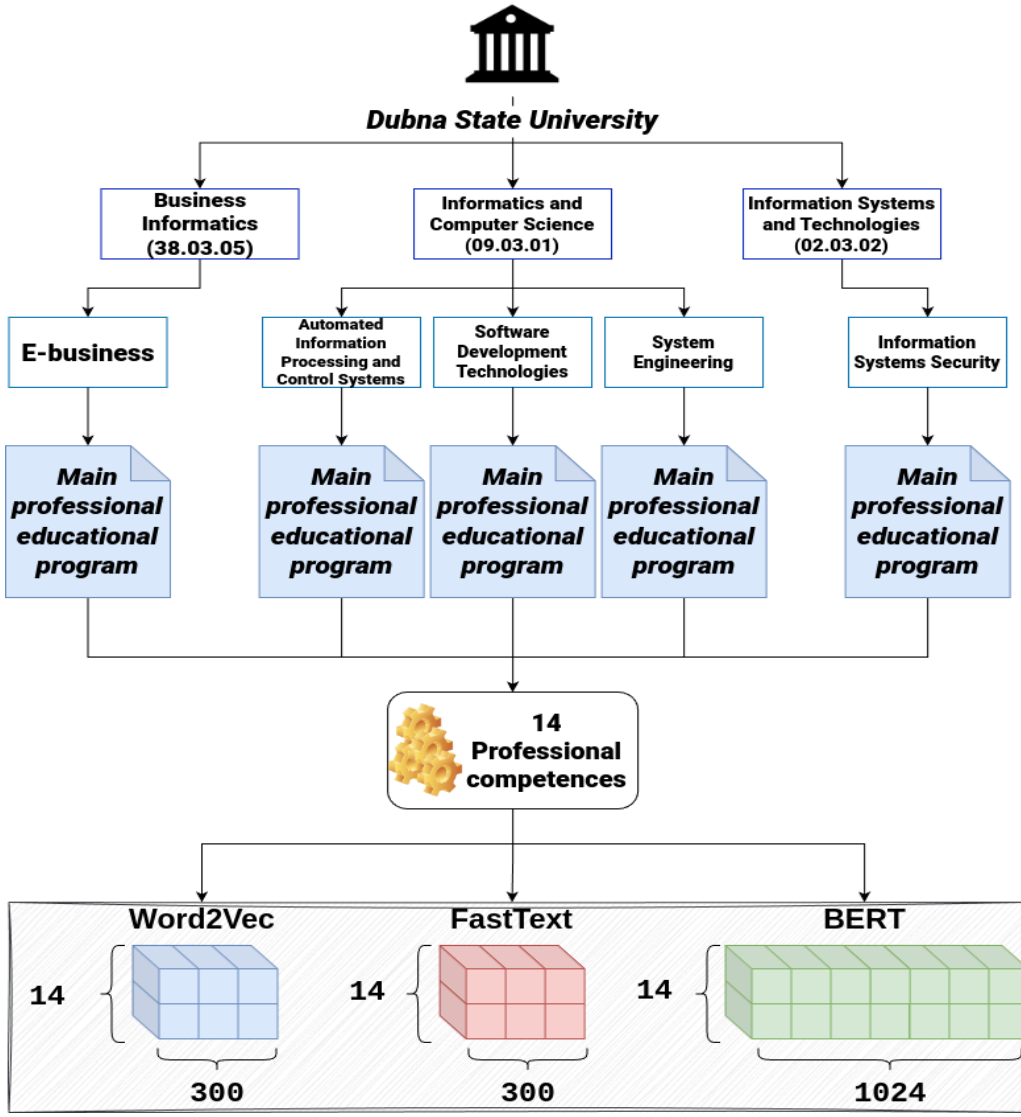
$$\vec{V}(S)=\frac{\sum_{i=1}^{k} p_i * \vec{V}(w_i)}{\sum_{i=1}^{k} p_i}$$

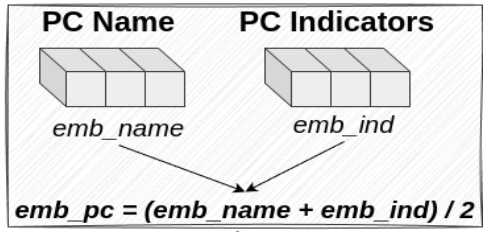
$$\cos(\vec{V}(w_1),\vec{V}(w_2))=\frac{\vec{V}(w_1)\times\vec{V}(w_2)}{\vec{V}(w_1)\cdot\vec{V}(w_2)}$$



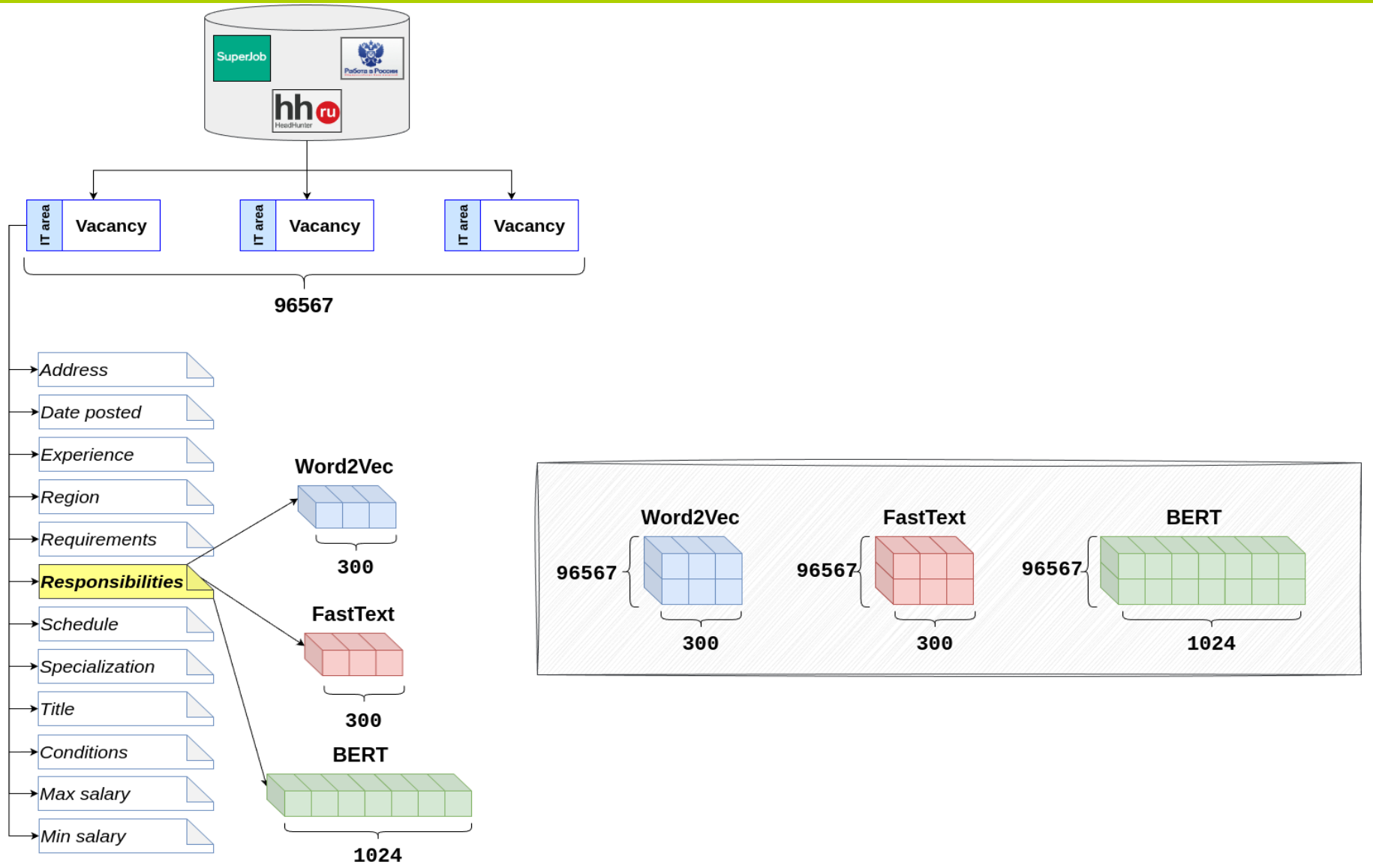
Input data. **Education system**



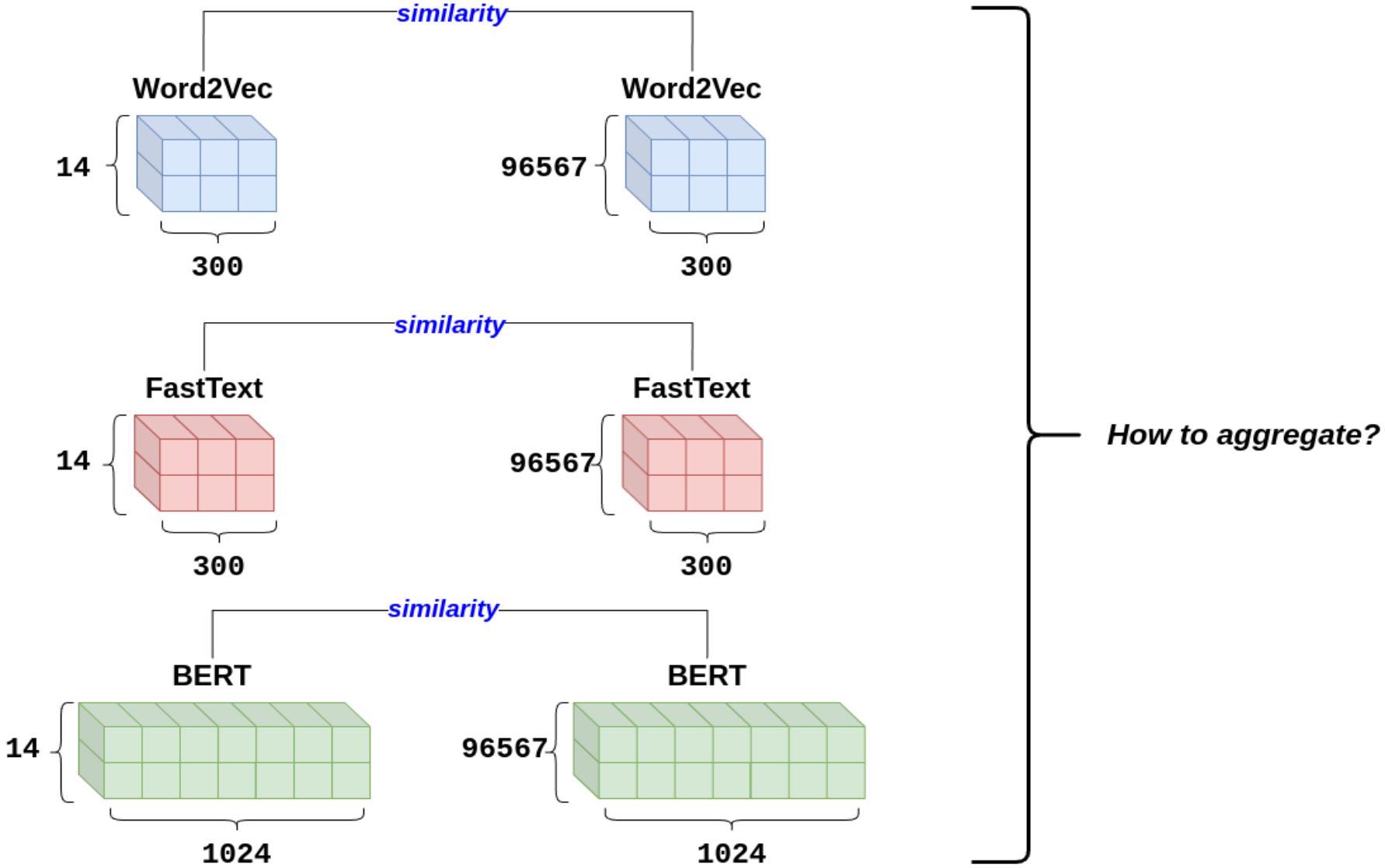




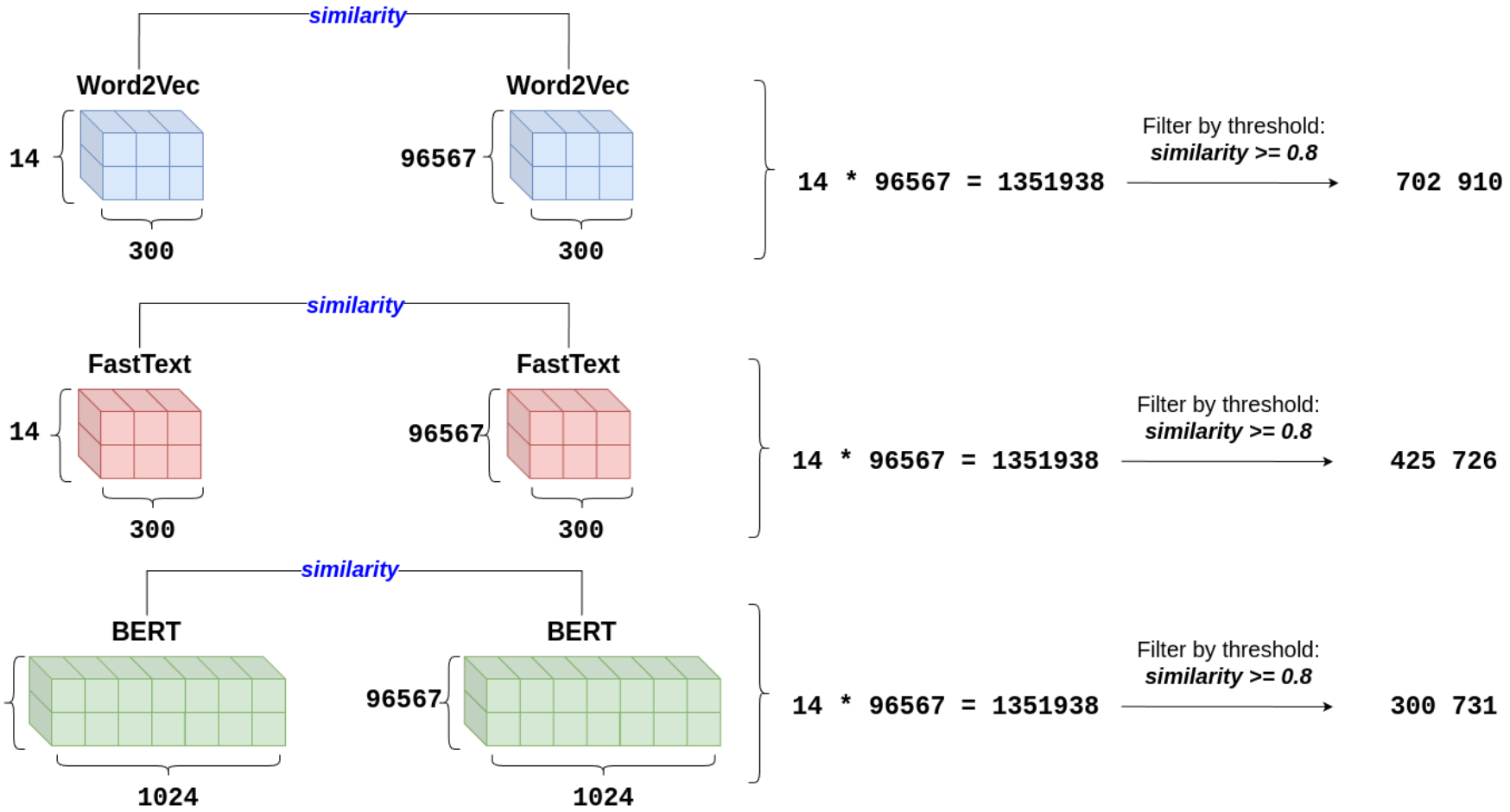
Input data. **Labor market**

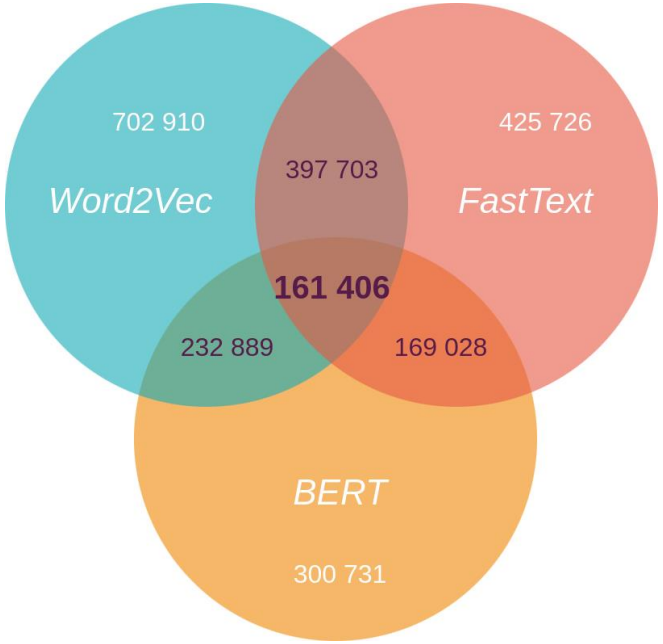


Search for aggregation methods



Data volume


similarity
Word2Vec
Word2Vec
14
96567
300
300
14 * 96567 = 1351938
Filter by threshold:
similarity >= 0.8
702 910
similarity
FastText
FastText
14
96567
300
300
14 * 96567 = 1351938
Filter by threshold:
similarity >= 0.8
425 726
similarity
BERT
BERT
14
96567
1024
1024
14 * 96567 = 1351938
Filter by threshold:
similarity >= 0.8
300 731


702 910
425 726
Word2Vec
397 703
FastText
161 406
232 889
169 028
BERT
300 731

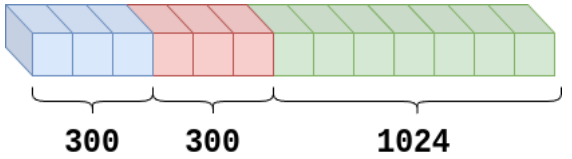
Evaluated aggregation methods

1. Averaging of similarities:

$$\frac{sim_{W2V}(v_i, c_j) + sim_{FT}(v_i, c_j) + sim_{BERT}(v_i, c_j)}{3}$$

2. Embeddings concatenation [19-21]:

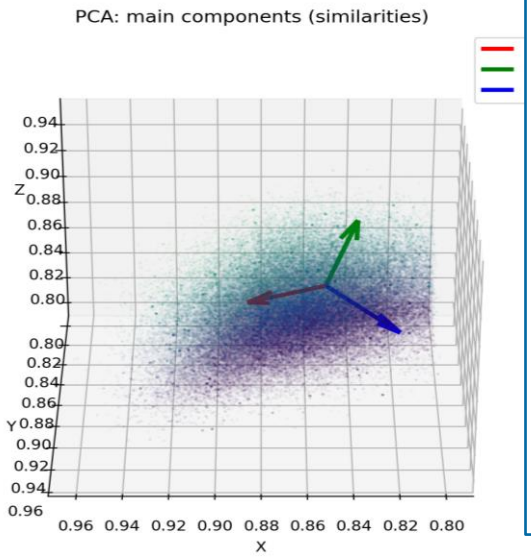
$$emb_{v_i} = emb_{W2V}(v_i) \cup emb_{FT}(v_i) \cup emb_{BERT}(v_i)$$
$$emb_{c_j} = emb_{W2V}(c_j) \cup emb_{FT}(c_j) \cup emb_{BERT}(c_j)$$

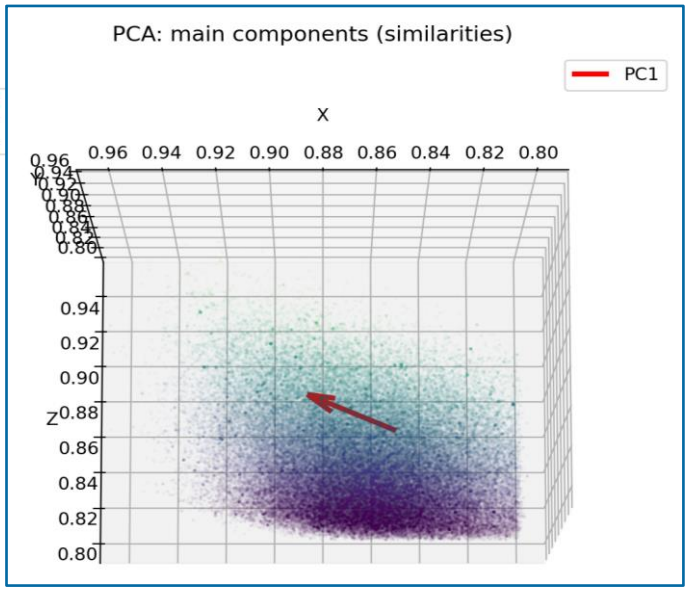


$$sim(emb_{v_i}, emb_{c_j})$$

3. Similarities aggregation using PCA

4. Similarities aggregation using SVD





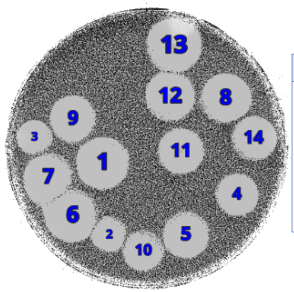
Used methods for consistency evaluation

1) Elements of graph analysis
2) Boxplots [22]
3) Calculation of the coefficient of variation (CV) [23, 24]
4) Correlation matrices [19, 25]
5) Calculation of the Kendall's concordance coefficient [19, 21].

Graph analysis

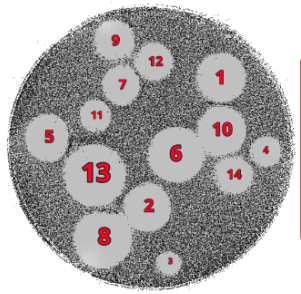
threshold:
cosine similarity >= 0.8

Word2Vec



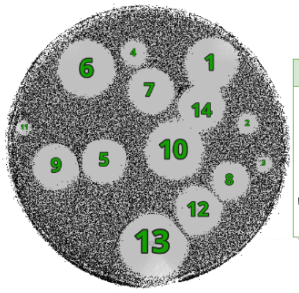
Metrics	
Nodes num	74 840
Edges num	702 856
Average degree	18,783
Weighted average degree	15,851

FastText

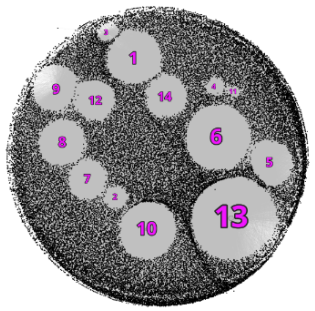


Metrics	
Nodes num	68 133
Edges num	425 693
Average degree	12,496
Weighted average degree	10,418

BERT

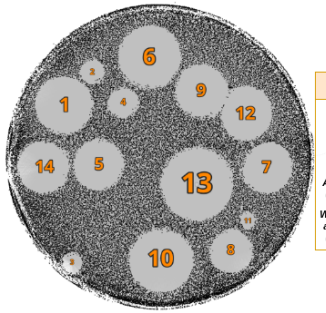


Metrics	
Nodes num	53 265
Edges num	300 731
Average degree	11,292
Weighted average degree	9,379



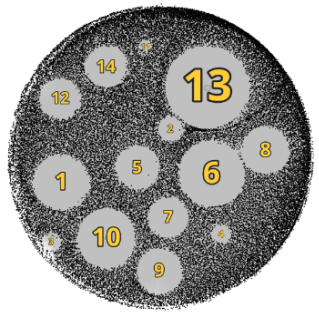
Metrics	
Nodes num	43 018
Edges num	161 406
Average degree	7,504
Weighted average degree	6,350

Average cosine similarity



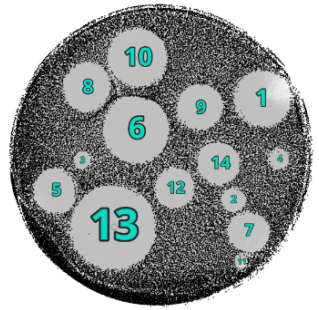
Metrics	
Nodes num	53 165
Edges num	299 455
Average degree	11,265
Weighted average degree	9,355

Vectors concatenation



Metrics	
Nodes num	43 018
Edges num	161 406
Average degree	7,504
Weighted average degree	6,384

PCA

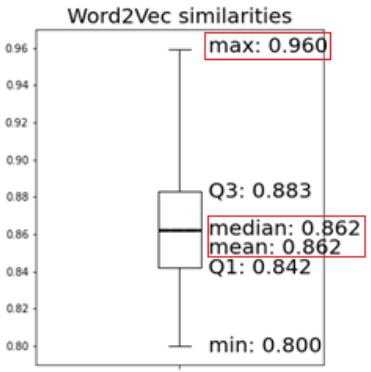


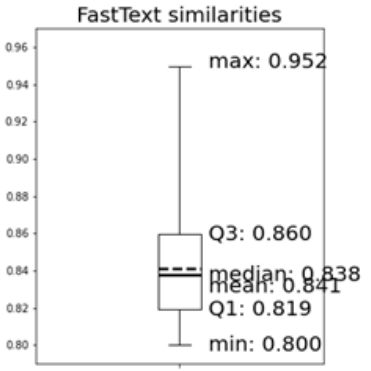
Metrics	
Nodes num	43 018
Edges num	161 406
Average degree	7,504
Weighted average degree	6,387

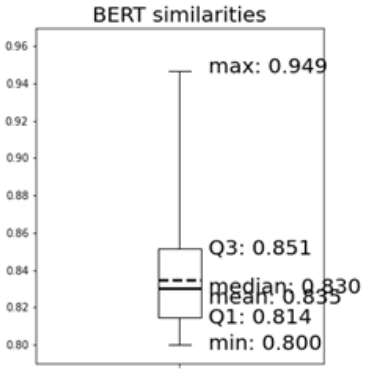
SVD

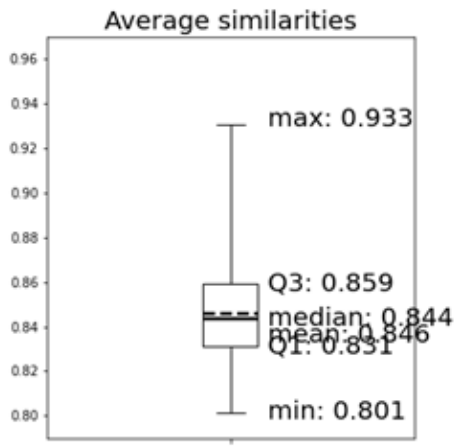
- All models and aggregates detect large, stable clusters → the main ‘centres of demand’ — i.e. frequently occurring or closely related competencies — are **identified by all methods, despite architectural differences**.
- **Word2Vec forms the densest and most connected graph**, indicating the model's tendency to associate more pairs as ‘similar.’
- **FastText and BERT are more selective**, especially BERT: it has fewer nodes and edges, indicating a more ‘sharp’ semantic distinction.
- **The similarities averaging, PCA, and SVD methods produce almost identical graphs, which indicates the consistency of these methods**.

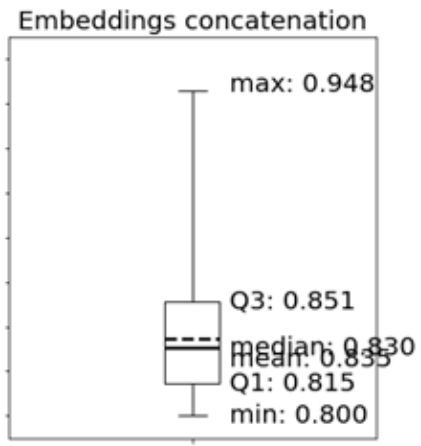
Boxplots

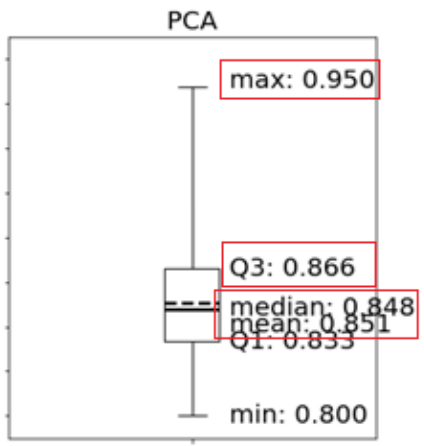


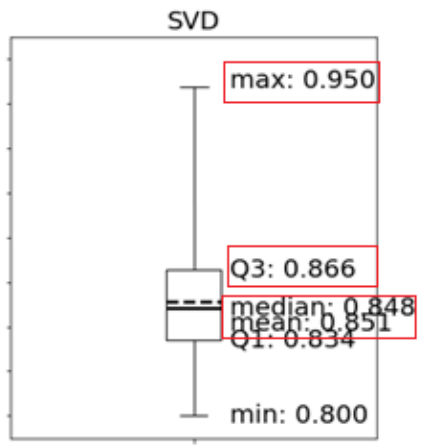












- Models:
  - Highest median and maximum → **Word2Vec tends to overestimate the closeness between concepts.** *Word2Vec generates more 'generous' similarities, possibly due to the limited context and simplicity of the model.*

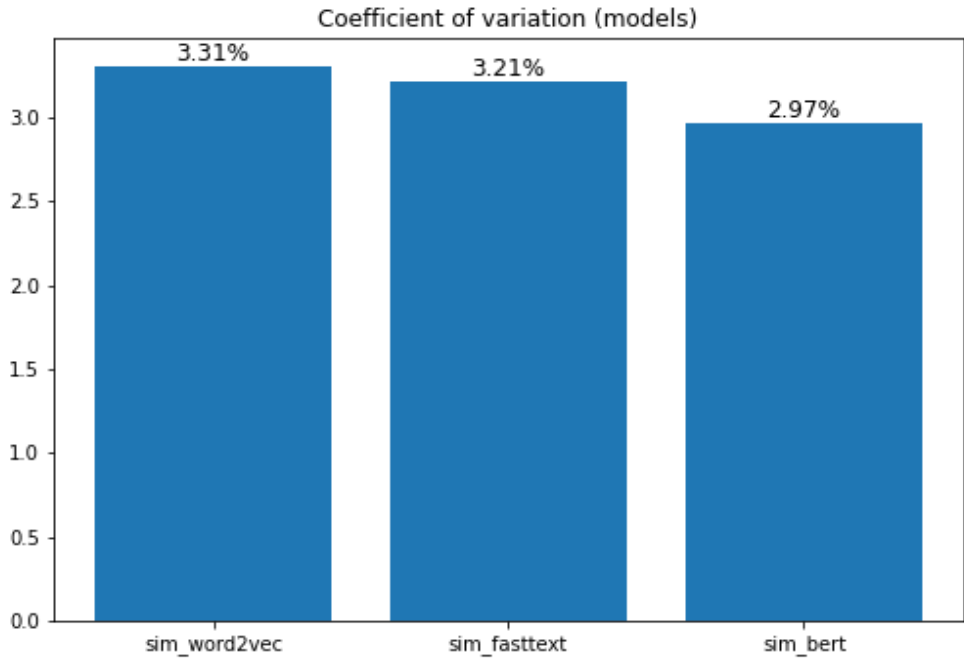
- Aggregation methods:
  - Low IQR →smooth out differences between models and produce stable estimates, creating a more **homogeneous space of similarity**.
  - For most methods, the mean value is similar to the median, indicating **no heavy skewness**.
  - PCA and SVD have the same IQR, mean, median and maximum

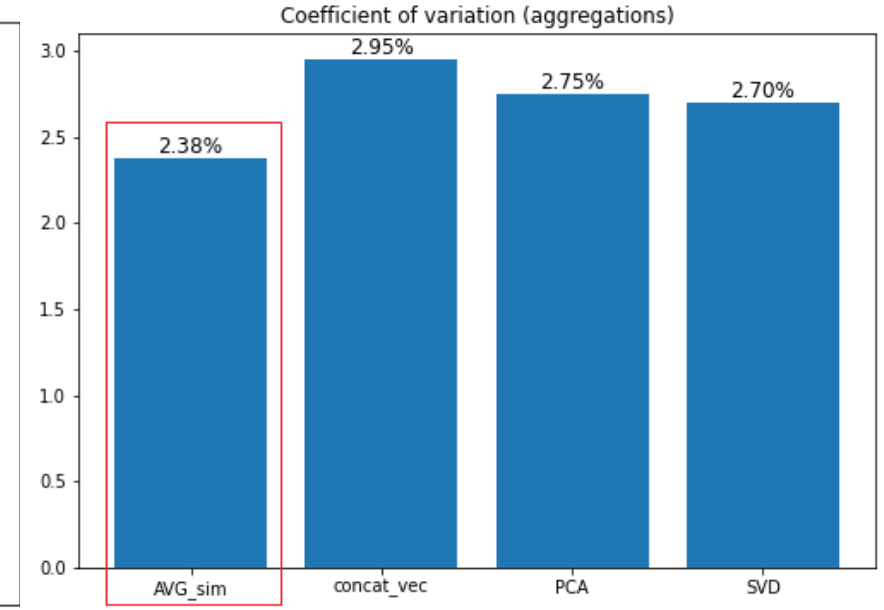
The coefficient of variation (CV) – a relative measure of the spread of the data

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

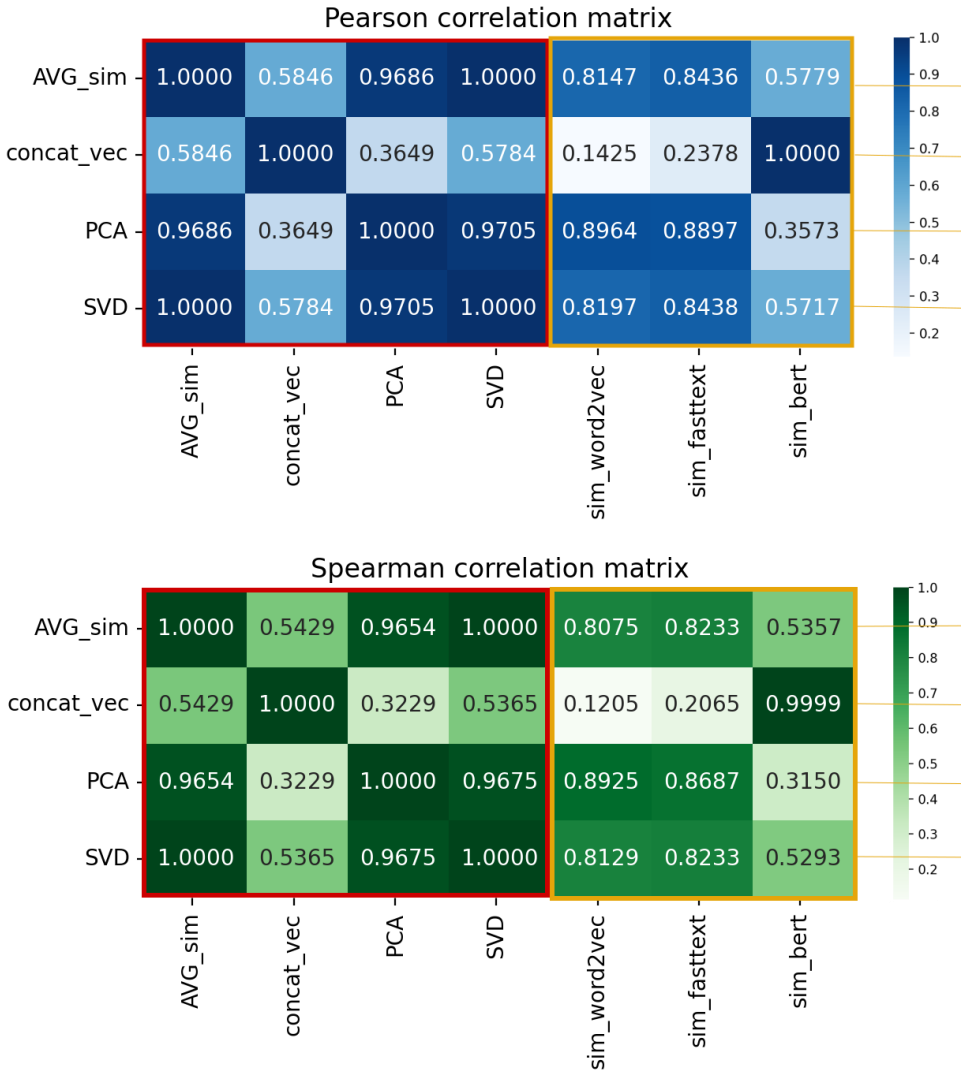




All coefficients of variation are close to zero, which indicates **sufficient homogeneity of the obtained proximity values** (both for the base models and aggregations).

The lowest coefficient of variation is for the averaging method (Average similarity) (2.38%), i.e. the aggregated results obtained by this method are the most homogeneous.

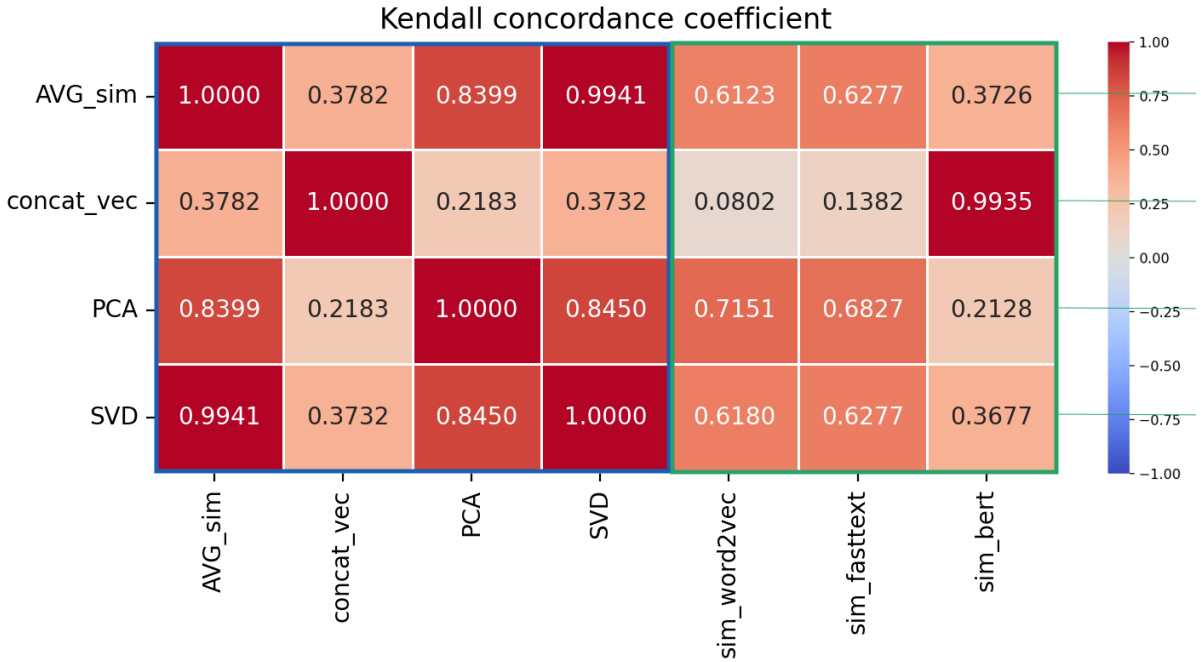
Correlation matrices

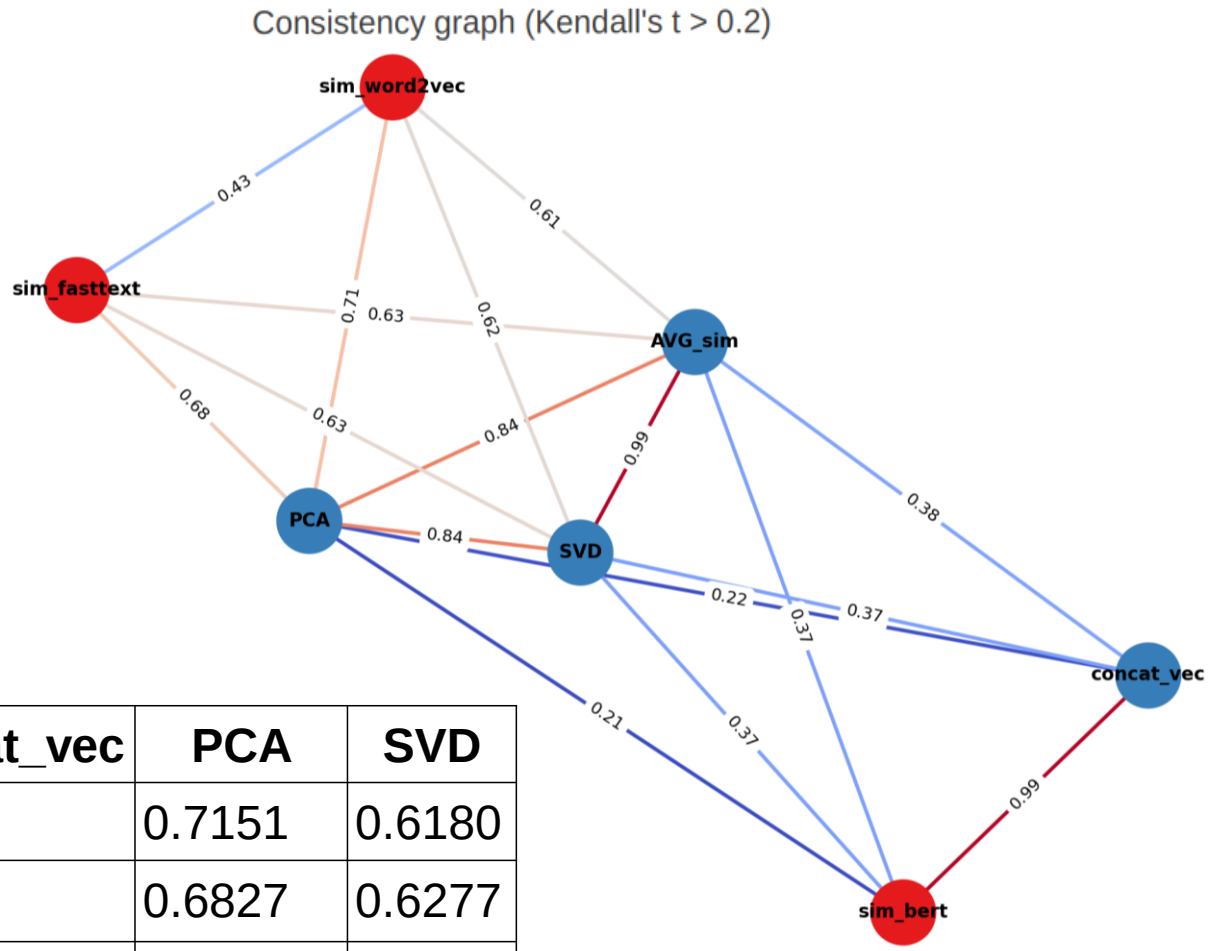


Aggregation method	Word2Vec (Pearson / Spearman)	FastText (Pearson / Spearman)	BERT (Pearson / Spearman)	MEAN (Pearson / Spearman)
Average similarity	0.81 / 0.81	0.84 / 0.82	0.58 / 0.54	**0.75 / 0.72**
Embeddings concatenation	0.14 / 0.12	0.24 / 0.21	1.0 / 0.99	0.46 / 0.44
PCA	0.90 / 0.89	0.89 / 0.87	0.36 / 0.32	0.72 / 0.69
SVD	0.82 / 0.81	0.84 / 0.82	0.57 / 0.53	**0.75 / 0.72**

- Aggregation methods correlate very well with Word2Vec and FastText
- **BERT shows low correlation with all aggregation methods**
- The almost complete match between the embedding concatenation method and the BERT model suggests that **BERT dominates the concatenation vector**, probably due to scale or weights. *Concatenating vectors without normalisation/weighting can result in one model (BERT) completely dominating the representation.*

Kendall's concordance coefficient (τ)



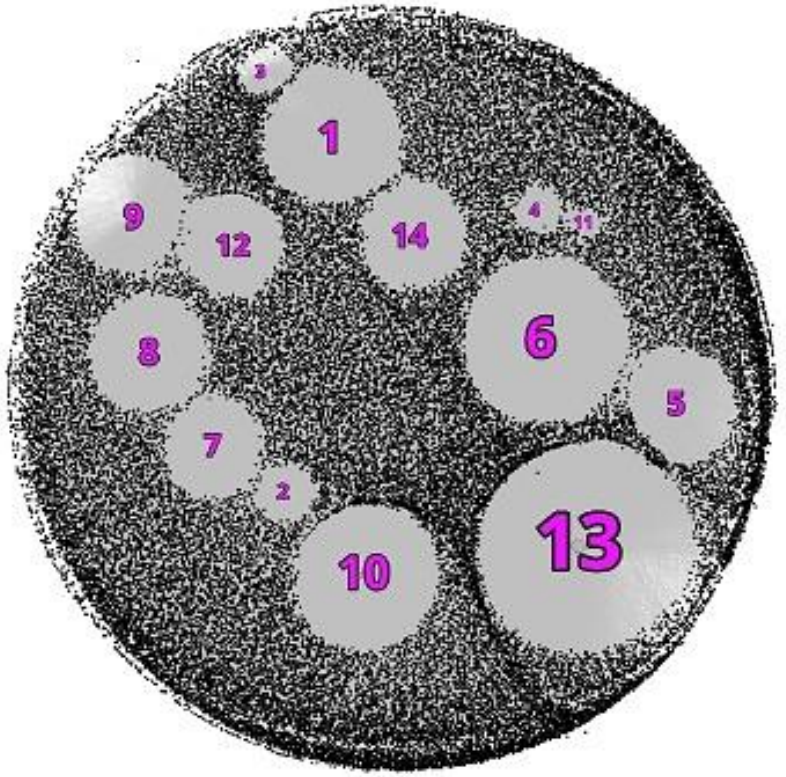


Method	Kendall's τ
AVG & SVD	**0.9941**
PCA & SVD	0.8450
PCA & AVG	0.8399

Method / Model	AVG_sim	Concat_vec	PCA	SVD
Word2Vec	0.6123	0.0802	0.7151	0.6180
FastText	0.6277	0.1382	0.6827	0.6277
BERT	0.3726	0.9935	0.2128	0.3677
MEAN	**0.5375**	0.4040	0.5369	**0.5378**

Findings

- Across all examined characteristics, the method based on concatenating the embeddings from the three models followed by computing semantic similarity between the resulting vectors (*concat_vec*) stands out. The low correlation of this method with the baseline models Word2Vec and FastText, combined with an excessively high correlation with BERT, suggests **a bias in favor of BERT**. *This effect is likely due to the significantly larger dimensionality of BERT's embeddings (1024 vs. 300 for Word2Vec and FastText).*
- **The most similar results are produced by methods that average the similarity scores** obtained from the baseline models Word2Vec, FastText, and BERT—specifically, the calculation of the average similarity (AVG_sim), as well as the aggregation of similarity values using PCA and SVD.
- All coefficients of variation (CV) are close to zero, indicating a sufficient degree of homogeneity in the obtained similarity values — both for the baseline models and the aggregation methods. **The lowest CV is observed for the AVG_sim method (2.38%), suggesting that the aggregated results produced by this approach are the most consistent**.
- A comparison of the Pearson and Spearman correlation matrices also reveals that **the AVG_sim method demonstrates the highest average similarity to the baseline models**, with correlation coefficients of 0.7454 (Pearson) and 0.7222 (Spearman), respectively.
- **The highest average Kendall's concordance coefficients with the baseline models are observed for the AVG_sim and SVD methods** (0.5375 and 0.5378, respectively), indicating that in more than half of the cases, the expert-like ranking of "Vacancy–Competence" pairs produced by these aggregation methods aligns with that of the individual baseline models.
- **The final aggregation strategy for combining similarity estimates from the selected embedding models should be based on the average similarity method (AVG_sim), as it demonstrates the lowest coefficient of variation and the highest correlation with the baseline models — thus most closely approximating their "consensus judgment."**



Metrics	
Nodes num	43 018
Edges num	161 406
Average degree	7,504
Weighted average degree	6,350

Average cosine similarity

Conclusion

This study investigates various methods for integrating the outputs generated by multiple embedding models, aiming to combine their individual strengths into a unified representation. In addition to exploring integration techniques, **the research focuses on evaluating the consistency of the resulting aggregated embeddings** both internally — by comparing different aggregation methods — and externally — by benchmarking against the baseline embedding models used independently.

Through analysis, **the method based on averaging similarity scores across models was identified as the most effective aggregation strategy**. This approach demonstrated superior stability, coherence, and alignment with the baseline models' semantic assessments, thereby providing a reliable consensus representation.

The insights and results obtained from this work have practical implications for enhancing the performance and robustness of semantic similarity models applied to textual data. In particular, the use of embedding ensembles can lead to more accurate and stable semantic comparisons, facilitating applications such as information retrieval, natural language understanding, and text classification.

In future research, the following directions are envisaged:

- The methodological toolkit for embedding aggregation is planned to be expanded through **the incorporation of trainable approaches**, particularly mechanisms based on attention architectures (attention-based fusion).
- An empirical evaluation of the effectiveness of the resulting aggregated representations is intended to be conducted across **applied downstream tasks**, including classification, semantic search, clustering, and topic modeling.

References

1. Lazarev Vladimir Nikolaevich, Pirogova Elena Vladimirovna, Zabolotnikova Maria Vladimirovna **Interaction of the educational services market and the labor market: problems and prospects** // Vestnik UlGTU. 2018. №1 (81). URL: https://cyberleninka.ru/article/n/vzaimodeystvie-rynka-obrazovatelnyh-uslug-i-rynka-truda-problemy-i-perspektivy (дата обращения: 04.07.2025).

2. Fedolyak V. S. **Mismatch of the educational services market with the requirements of the labor market: reasons and ways to overcome** // Vocational Guidance. 2018. №2. URL: https://cyberleninka.ru/article/n/nesootvetstvie-rynka-obrazovatelnyh-uslug-trebovaniyam-rynka-truda-prichiny-i-sposoby-preodoleniya (дата обращения: 04.07.2025).

3. Carla Varona Cervantes, Russell Cooper, **Labor market implications of education mismatch**, European Economic Review, Volume 148, 2022, p. 104179, ISSN 0014-2921, https://doi.org/10.1016/j.euroecorev.2022.104179.

4. **Monitoring of labor market needs in university graduates based on data-intensive analytics** / P. V. Zrelov, 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. -C. 124-131.

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

6. **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. DLCP2022. – Pp. 028. – DOI: https://doi.org/10.22323/1.429.0028.

7. Belov, S.D., Zrelov, 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.

8. Belov, S., Ilina, 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.

9. **MEET: A Method for Embeddings Evaluation for Taxonomic Data**/ L. Malandri, F.Mercori, M. Mezzanzanica, N. Nobani// 2020 International Conference on Data Mining Workshops (ICDMW), 2020. –Pp.31-38.–DOI:10.1109/ICDMW51313.2020.00014

10. **An AI-based open recommender system for personalized labor market driven education** / M. Tavakoli, A. Faraji, J.Vrolijk, M.Molavi[et al.] //Advanced Engineering Informatics, 2022. –Vol. 52. –Pp.101508.–DOI: https://doi.org/10.1016/j.aei.2021.101508

11. **OntoJob: Automated Ontology Learning from Labor Market Data** / J. Vrolijk, S. T. Mol, C. Weber, M. Tavakoli[et al.] // 2022 IEEE 16th International Conference on Semantic Computing (ICSC). –2022 Pp. 195-200.–DOI:10.1109/ICSC52841.2022.00040

12. WowczkoI.A. **Skills and Vacancy Analysis with Data Mining Techniques**. –Informatics. –2015. –Vol. 2 (4). –Pp. 31-49. –DOI:10.3390/informatics2040031.

References

13. **Classifying online Job Advertisements through Machine Learning** / R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica//Future Generation Computer Systems. –Vol.86, Issue C. –2018. –Pp.319-328.

14. SparreboomT**., Labour market information and analysis systems** // Perspectives on labour economics for development, Geneva: ILO, 2013. –Pp. 255-282.

15. D. A. Tamburri, W. -J. V. D. Heuvel and M. Garriga, "**DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching**," 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 2020, pp. 391-394, doi: 10.1109/IRI49571.2020.00063.

16. Nikita Matkin and Aleksei Smirnov and Mikhail Usanin and Egor Ivanov and Kirill Sobyanin and Sofiia Paklina and Petr Parshakov. **Comparative Analysis of Encoder-Based NER and Large Language Models for Skill Extraction from Russian Job Vacancies**. 2024. https://arxiv.org/abs/2407.19816

17. Lyubov Aleksandrovna Komarova, Alexey Mikhailovich Kolosov, Vladimir Igorevich Soloviev **Matching vector representations of vacancies and resumes using large language models** // International Journal of Open Information Technologies. 2025. №2. URL: https://cyberleninka.ru/article/n/sopostavlenie-vektornyh-predstavleniy-vakansiy-i-rezyume-s-ispolzovaniem-bolshih-yazykovyh-modeley.

18. Frank D. Zamora-Reina and Felipe Bravo-Marquez and Dominik Schlechtweg. **LSCDiscovery: A shared task on semantic change discovery and detection in Spanish**. 2022. https://arxiv.org/abs/2205.06691

19. 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.

20. Ruan, William. **Word Embeddings using Canonical Correlation Analysis**. Faculté des sciences, Université catholique de Louvain, 2022. Prom. : Pircalabelu, Eugen. http://hdl.handle.net/2078.1/thesis:38059

21. Mingze Dong, David Su, Harriet Kluger, Rong Fan, Yuval Kluger, **SIMVI 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

22. Ding, Xiaoyu and Yang, Yihong and Stein, Elliot A. and Ross, Thomas J**., Combining Multiple Resting-State fMRI Features during Classification: Optimized Frameworks and Their Application to Nicotine Addiction, Frontiers in Human Neuroscience**, Volume 11, 2017, DOI:10.3389/fnhum.2017.00362.

23. 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.

24. 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

25. 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⟩

Thank you for your attention!

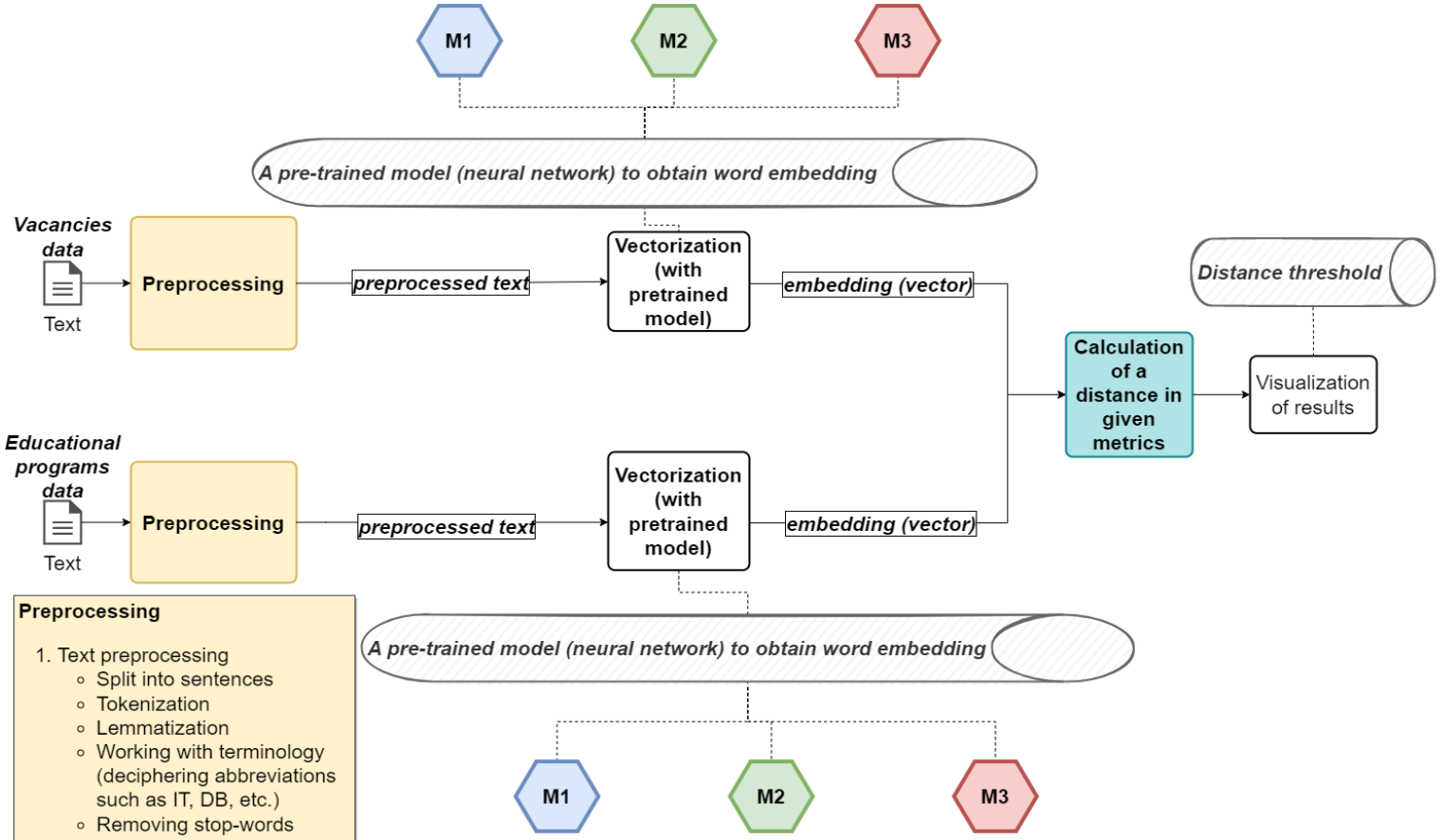
Slides reserve

```
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). "ЦЭК" стабильно...
responsibilities	: - Создать сайт с нуля на популярном движке, "натянуть дизайн", 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

PCs of Dubna State University

ID of PC	Title of PC	Indicators of PC achievements
…		
4	Способен проектировать и разрабатывать компоненты корпоративных информационных систем и информационных систем электронного бизнеса	Проектирует компоненты информационных систем электронного бизнеса; Разрабатывает компоненты информационных систем
6	Способен выполнять проектную деятельность по разработке и созданию (модификации) ИС, разработке технико-экономического обоснования проектов для улучшения бизнес-процессов и ИТ-инфраструктуры предприятия	Реализует проекты по разработке и модификации информационных систем; Разрабатывает технико-экономическое обоснование проектов по улучшению бизнес-процессов и ИТ-инфраструктуры предприятия
…		
13	Способен проектировать и создавать программное обеспечение, соответствующее требованиям заказчика, включая разработку программного интерфейса…	Осуществляет проектирование и разработку программных модулей и компонентов, соответствующих требованиям заказчика, включая разработку программного интерфейса…



Preprocessing

1. Text preprocessing
   - Split into sentences
   - Tokenization
   - Lemmatization
   - Working with terminology (deciphering abbreviations such as IT, DB, etc.)
   - Removing stop-words
   - (*optional*) Tagging
2. (*optional*) Calculation of TF-IDF word weights
3. (*optional*) Identifying the most meaningful text from the three levels of PC mastery ("Know", "Can", "Possess") (for PC PRUE) using the Principal Component Analysis method

Distance metrics

1. Cosine distance

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