

COMPARATIVE ANALYSIS AND EVALUATION OF THE CONSISTENCY OF METHODS FOR INTEGRATING SEMANTIC REPRESENTATIONS OF TEXTS

A. Ilina^{1, *}, *P. Zrelov*^{1, **}

¹ Joint Institute for Nuclear Research, Dubna, Russia

We explore methods for aggregating the results of three embedding models to estimate semantic similarity of texts: cosine similarities averaging, vector concatenation, and selection of one of the three cosine similarities based on principal component analysis and singular value decomposition methods. Statistical analysis showed that averaging cosine proximity as an aggregation method demonstrates the best results. The obtained results can be used to improve the robustness of models for semantic proximity evaluation of texts when using embedding ensembles.

Исследуются методы агрегации результатов трех эмбединговых моделей для оценки семантической близости текстов: усреднение косинусных близостей, конкатенация векторов, а также выбор одной из трех косинусных близостей на основе методов анализа главных компонент и сингулярного разложения. Статистический анализ показал, что усреднение косинусных близостей в качестве метода агрегации показывает наилучшие результаты. Полученные данные могут быть использованы для повышения устойчивости моделей оценки семантической близости текстов с помощью ансамблей эмбедингов.

PACS: 89.20.Ff

* E-mail: annailina@jinr.ru

** E-mail: zrelov@jinr.ru