

Early Slavic word embeddings

Nilo Pedrazzini

1 Description

Link: <https://zenodo.org/records/8414137>

The Zenodo repository contains word embeddings trained on the lemmatised TOROT Treebank (Eckhoff & Berdičevskis, 2015), using Word2Vec (Mikolov et al., 2013) as implemented in the Gensim library (Řehůřek & Sojka, 2010).

The following parameters were used:

- `sg = True`
- `min_count = <1,3,5>`
- `window = <3,5>`
- `vector_size = <100,200,300>`
- `epochs = 5`

One model was trained for each combination of the parameters enclosed in angled brackets (`< >`).

The release contains both the full models (`.model`) and the plain vector files (`_vectors.txt`). The models are named according to the parameters they were trained with.

2 Content and file sizes

- **TOROT_lemmat_models.zip**

– *Minimum frequency: 1*

* TOROT_lemmat_min1_win3_dims100_ep5.model	15.1 MB
* TOROT_lemmat_min1_win3_dims100_ep5_vectors.txt	22.2 MB
* TOROT_lemmat_min1_win3_dims200_ep5.model	29.5 MB
* TOROT_lemmat_min1_win3_dims200_ep5_vectors.txt	44.4 MB
* TOROT_lemmat_min1_win3_dims300_ep5.model	43.9 MB
* TOROT_lemmat_min1_win3_dims300_ep5_vectors.txt	66.9 MB
* TOROT_lemmat_min1_win5_dims100_ep5.model	15.1 MB

* TOROT_lemmat_min1_win5_dims100_ep5_vectors.txt	22.0 MB
* TOROT_lemmat_min1_win5_dims200_ep5.model	29.5 MB
* TOROT_lemmat_min1_win5_dims200_ep5_vectors.txt	44.0 MB
* TOROT_lemmat_min1_win5_dims300_ep5.model	43.9 MB
* TOROT_lemmat_min1_win5_dims300_ep5_vectors.txt	66.6 MB
– <i>Minimum frequency: 3</i>	
* TOROT_lemmat_min3_win3_dims100_ep5.model	6.5 MB
* TOROT_lemmat_min3_win3_dims100_ep5_vectors.txt	9.2 MB
* TOROT_lemmat_min3_win3_dims200_ep5.model	12.6 MB
* TOROT_lemmat_min3_win3_dims200_ep5_vectors.txt	18.6 MB
* TOROT_lemmat_min3_win3_dims300_ep5.model	18.8 MB
* TOROT_lemmat_min3_win3_dims300_ep5_vectors.txt	28.1 MB
* TOROT_lemmat_min3_win5_dims100_ep5.model	6.5 MB
* TOROT_lemmat_min3_win5_dims100_ep5_vectors.txt	9.2 MB
* TOROT_lemmat_min3_win5_dims200_ep5.model	12.6 MB
* TOROT_lemmat_min3_win5_dims200_ep5_vectors.txt	18.5 MB
* TOROT_lemmat_min3_win5_dims300_ep5.model	18.8 MB
* TOROT_lemmat_min3_win5_dims300_ep5_vectors.txt	28.0 MB
– <i>Minimum frequency: 5</i>	
* TOROT_lemmat_min5_win3_dims100_ep5.model	4.5 MB
* TOROT_lemmat_min5_win3_dims100_ep5_vectors.txt	6.3 MB
* TOROT_lemmat_min5_win3_dims200_ep5.model	8.7 MB
* TOROT_lemmat_min5_win3_dims200_ep5_vectors.txt	12.7 MB
* TOROT_lemmat_min5_win3_dims300_ep5.model	13.0 MB
* TOROT_lemmat_min5_win3_dims300_ep5_vectors.txt	19.2 MB
* TOROT_lemmat_min5_win5_dims100_ep5.model	4.5 MB
* TOROT_lemmat_min5_win5_dims100_ep5_vectors.txt	6.3 MB
* TOROT_lemmat_min5_win5_dims200_ep5.model	8.7 MB
* TOROT_lemmat_min5_win5_dims200_ep5_vectors.txt	12.7 MB
* TOROT_lemmat_min5_win5_dims300_ep5.model	13.0 MB
* TOROT_lemmat_min5_win5_dims300_ep5_vectors.txt	19.2 MB

3 Disclaimer

Note that these are the result of very preliminary experiments and no systematic evaluation of their quality was carried out, so use with caution.

References

- Eckhoff, Hanne M. & Aleksandrs Berdičevskis. 2015. Linguistics vs. Digital Editions: The TromsøOld Russian and OCS Treebank. *Scripta & E-Scripta* 14-15. 9–25.
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. doi:10.48550/ARXIV.1301.3781. <https://arxiv.org/abs/1301.3781>.
- Řehůřek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. <http://is.muni.cz/publication/884893/en>.