Word2Vec implementation for Russian text

Olha Kaminska
Master student
University of Tartu
Institute of Computer Science

July 18, 2018

Abstract

In our time, the task of human language processing is one of the most important branch in the field of Machine Learning. Many tasks require working with text data, and a lot of information is stored in text format. But Machine Learning algorithms work with numbers, not letters, so developing a correct system for converting words to numbers is an important and relevant task.

In this work, the implementation of the algorithm for converting words into vectors, called Word2Vec, is considered. The algorithm is implemented for the Russian text, because, unlike English, for this language exist not so many optimal solutions in the field of Natural Language Processing. The implemented algorithm turns the given word into a 300-dimensional vector, such that the words close in meaning have close coordinates. In general, the described algorithm can be applied to other languages, which can provide enough input text data for processing and vocabulary building.

Work on this project can be divided on the next main steps: collect Russian text data, implement Word2Vec algorithm for Russian text, test obtained results. To check how well implementation works, it was used in movie reviews classification task. All words from train and test data were transformed in vectors, using implemented Word2Vec, then Logistic Regression model with default parameters was applied.

KEYWORDS: WORD2VEC, NATURAL LANGUAGE PROCESSING, WORD EMBEDDING, CLASSIFICATION, RUSSIAN TEXT ANALYSIS