СЕМАНТИЧЕСКОЕ КОДИРОВАНИЕ СЛОВ С ПОМОЩЬЮ WORD2VEC

Крощенко Александр Александрович

Преподаватель кафедры прикладной математики и информатики
УО «Брестский государственный университет имени А.С. Пушкина»

Задача семантического кодирования приобрела особую важность с развитием поисковых систем. Актуальность подобных технологий связана в первую очередь с возможностью осуществления поиска в больших по объему базах. При этом особое значение имеет не столько нахождение идентичных слов, сколько осуществление поиска близких по некоторой семантической метрике слов.

В силу того, что слова в словаре некоторого языка почти всегда отличаются по длине, реализация какой-либо задачи сравнения слов существенно усложняется. Приведение же каждого слова словаря к вектору заданной размерности позволяет осуществлять сравнение искомого и проверяемого слов непосредственно путем вычисления любой (например, евклидовой метрики). Такая технология позволяет упростить задачи поиска.

Одним из методов семантического кодирования, широко применяемым на практике, является word2vec. Этот подход был предложен Миколовым в 2013 году [1].

Word2vec позволяет осуществлять семантический анализ текста с выделением наиболее близких по смыслу слов. Существует два варианта метода word2vec, отличающихся политикой участия контекста. Под контекстом в данном случае понимается совокупность слов (слева и справа), окружающая целевое слово, взятая в пределах определенного окна.

Первый вариант, называемый skip-grams, базируется на обучении нейросетевой модели, которая осуществляет формирование контекста на основе одного целевого слова, подаваемого на вход модели.

Второй вариант, называемый CBOW (Continuous Bag of Words), использует нейронную сеть для получения целевого слова на основе подаваемого контекста.

В результате применения word2vec обучается искусственная нейронная сеть, которая осуществляет отображение слова, записанного в виде унитарного кода в пространство меньшей размерности, которое впоследствии используется для оценки семантической близости слов.

Для иллюстрации работы метода word2vec приведем пример двумерной визуализации редуцированных кодов слов, полученной нами для выборки

из 100.000 англоязычных документов википедии и общего размера словаря 50.000 слов. В данном эксперименте использовалась упрощенная архитектура skip-grams, включающая 50.000 входных нейронов, соответствующих целевому слову, 300 скрытых и 50.000 выходных нейронов, соответствующих контекстному слову.

При обучении были сформированы пары слов (целевое_слово, контекстное слово), которые подавались на нейронную сеть мини-батчами по 128 пар в каждом. После обучения к редуцированным кодам слов был применен алгоритм t-SNE [2] для уменьшения размерности данных. Фрагмент полученной двумерной карты семантического сходства изображен на рисунке ниже.

Карта семантического сходства

Приведенный рисунок иллюстрирует тот факт, что с помощью параметров обученной нейронной сети можно осуществлять поиск близких в семантическом плане слов, например, слова **lake**, **river**, **sea** и **water** попадают в одну группу, а слова **album**, **record**, **song** – в другую.

СПИСОК ЛИТЕРАТУРЫ

1. Mikolov, T. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv [Web-resource]. – 2013. – Mode of access: https://arxiv.org/pdf/1301.3781.pdf. – Date of access: 12.12.2017.

2. Van der Maaten, L. Visualizing High-Dimensional Data Using t-SNE /

L.J.P van der Maaten, G.E. Hinton // Journal of Machine Learning Research. – Volume 9. – 2008. – P. 2579–2605.