Д. А. Хвесюк
(ГГУ имени Ф. Скорины, Гомель)
Науч. рук. **Н. А. Аксёнова**, ст. преподаватель

ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ ДОКУМЕНТОВ С ПОМОЩЬЮ МОДЕЛИ DOC2VEC

В 2013 году была опубликована статься Efficient Estimation of Word Representations in Vector Space, которая описывала метод отображения слов в векторном пространстве – word2vec. До этого в сфере обработки естественного языка использовались методы кодирования слов не предполагающие наличия в них смысла. Однако после выпуска и реализации модели во всех популярных библиотеках машинного обучение, данный подход стал повсеместно использоваться на этапе трансформации данных, а значит мог использоваться при решении любой из задач машинного обучения: классификация, регрессия и другие. Также данное преобразование помогало кодировать не только слова, но и различные товары, что позволяло успешно использовать модель в рекомендательных и поисковых системах. Однако объектом рекомендации или поиска не всегда является товар. Все чаще и чаще подобные системы пытаются внедрять и в информационные продукты: блоги, новостные сайты, агрегаторы статей, где важно выделить смысл не конкретного слова, а целого документа. Так было предложена модель doc2vec.

Чтобы понять как работает данная модель, обратимся к простейшей реализации модели word2vec, а именно к методу CBOW (Continuous bag of words). В данном подходе мы используем скользящее окно вокруг текущего слова, чтобы предсказать его по контексту – окружающим словам. Каждое слово представлено в виде вектора признаков.

После обучения эти векторы становятся векторами слов. Похожие вектора должны так отображаться в пространстве, что значение метрики расстояния между ними будет минимальным.

Чтобы данная модель начала работать не со словами, а с документами, в процесс обучение надо включить еще один компонент – уникальный вектор документа, который также будет считаться контекстом конкретного слова. При обучении векторов слов также обучается и данный вектор, который в конце обучения и будет содержать нужное числовое представление.

Описанная выше модель называется Distributed Memory version of Paragraph Vector (PV-DM). Существует также модель Distributed Bag of Words version of Paragraph Vector (PV-DBOW), которая схожа с другим методом построения векторных представлений – skip-gram. Данная модель быстрее обучается, однако заметно теряет в качестве.

Так как модели являются модификациям word2vec, к ним применимы различные улучшения: Glove, Fast-Text. Однако модель имеет и свои методы улучшения. Например, добавление к тексту тегов или тем, которые также участвуют в обучении. Это позволяет использовать не только локальный контекст слов, но и контекст тем самих документов.