УДК 004.92

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПОИСКА ПОХОЖИХ ТЕКСТОВЫХ ДОКУМЕНТОВ

© Чеховских И.В., Симонова Е.В.

e-mail: chekhovskikh@mail.ru

Самарский национальный исследовательский университет имени академика С.П. Королёва г.Самара, Российская Федерация

Человек тратит много сил и времени на поиск интересующего текстового документа. Для автоматизации данного процесса необходимо настроить подбор текстовых документов на основе схожести информации в них, база данных при этом не требуется. Для подобных целей используются рекомендательные системы.

В большинстве таких систем необходимо выполнить предобработку текстового документа. Текст очищается от предлогов, союзов, знаков препинания и других слов, которые не должны участвовать в сравнении.

Программный продукт Doc2Vec предназначен для поиска текстовых документов на основе модели нейронной сети, задачей которой является реконструкция контекста слов. Преимуществом Doc2Vec является небольшая размерность векторов.

Принцип работы Doc2Vec можно описать следующим образом: максимизация косинусной близости для векторного представления текстовых документов, которые появляются в похожих контекстах, и, наоборот, её минимизация для документов, не встречающихся в похожих контекстах.

Для того чтобы использовать Doc2Vec, можно взять модель, обученную, например, на корпусе Википедии, или же обучить её самому. Недостаток уже готовой модели в том, что она может быть слишком общей и, соответственно, идентифицировать как близкие друг к другу слова, которые такими не являются.