РАЗРАБОТКА МЕТОДА ОЦЕНКИ СХОЖЕСТИ КОЛЛЕКЦИЙ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ ИХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ, ПОЛУЧЕННЫХ МЕТОДОМ DOC2VEC

А.Ю. Карпова, А.М. Ширыкалов
Томский политехнический университет
E-mail: ams28@tpu.ru

Введение

Автоматическая оценка схожести коллекций текстов представляет большой интерес, в том числе для исследователей, занимающихся изучением сообществ в социальных сетях. Мера схожести текстовых коллекций имеет множество применений, в том числе, построение систем автоматизированного сбора данных по определенной тематике, получение данных для создания различных визуализаций и так далее.

В данной работе предложен метод оценки сходства текстовых коллекций, основывающийся на сравнении их векторных представлений, полученных путем применения метода *doc2vec*, и приведен пример его применения для определения сходства текстового контента сообществ социальной сети.

Описание метода

Важным этапом подготовки текстовых данных к обработке является их векторизация. Метод векторизации текстовых данных *doc2vec*, являющийся развитием метода *word2vec* [1, 2], был впервые описан в 2014 году [3]. Обучение моделей машинного обучения на векторных представлениях, полученных этим методом, дает значительно лучшие результаты [3], чем обучение на векторах, полученных альтернативными методами, такими как *bag of words, tf-idf, SVM* и т. д.

Метод векторизации текста *doc2vec* основан на обучении нейронной сети с одним скрытым слоем на специальной синтетической задаче. Существует две версии метода *doc2vec*: *Distributed Memory* и *Distributed Bag of Words*, различие которых заключается в постановке этой задачи.

В варианте *Distributed Memory* модель обучается предсказывать вероятность появления слова в документе, имея на входе вектора нескольких других слов, предшествующих искомому, а также вектор самого документа, а в варианте *Distributed Bag of Words* – предсказывать вероятность появления случайных слов из документа по его вектору.

Описание метода оценки схожести текстовых коллекций

В качестве меры сходства двух векторов текстовых коллекций, полученных методом doc2vec, была использована мера cosine similarity (косинусное сходство), позволяющая получить значение в промежутке от -1 до 1, где 1 означает высокое сходство, 0 – отсутствие пересечений, -1 – противоположность.

Таким образом, был разработан следующий метод оценки схожести между текстовыми коллекциями. Предполагается, что на вход методу подается список коллекций, после чего на выходе получается оценка схожести для всех комбинаций их пар.

Разработанный метод состоит из следующих этапов:

1) Объединение текстов каждой коллекции в один документ (строку).
2) Добавление в корпус документов, содержащих различные тексты на русском языке. Так как в [3] было явно показано, что doc2vec работает лучше при увеличении объема обучающей выборки, для увеличения точности метода рекомендуется, чтобы общее количество документов было не менее миллиона. В ходе апробации метода использовались тексты в том числе из открытой базы новостей lenta.ru.
3) Удаление из каждого документа пунктуации, затем преобразование документов в списки строк, состоящие из их слов.
4) Обучение модели doc2vec на полученном корпусе.
5) Расчет косинусного сходства для каждой пары текстовых коллекций с использованием векторов, полученных для них в результате работы doc2vec.

Пример применения метода оценки схожести текстовых коллекций

Разработанный метод был опробован на выборке постов за один календарный год из 99 сообществ социальной сети «ВКонтакте» различной тематики, подобранных экспертным способом. Ниже, на рисунке 1, представлена визуализация в виде графа, на которой вершинами являются сообщества, а

наличие ребра между ними определяется значением полученной схожести. Вершины сообществ, имеющих схожую тематику, отмечены одинаковым цветом. Наличие ребра между вершинами означает, что схожесть соответствующих сообществ была больше 0.5, а его толщина прямо пропорциональна значению их схожести.

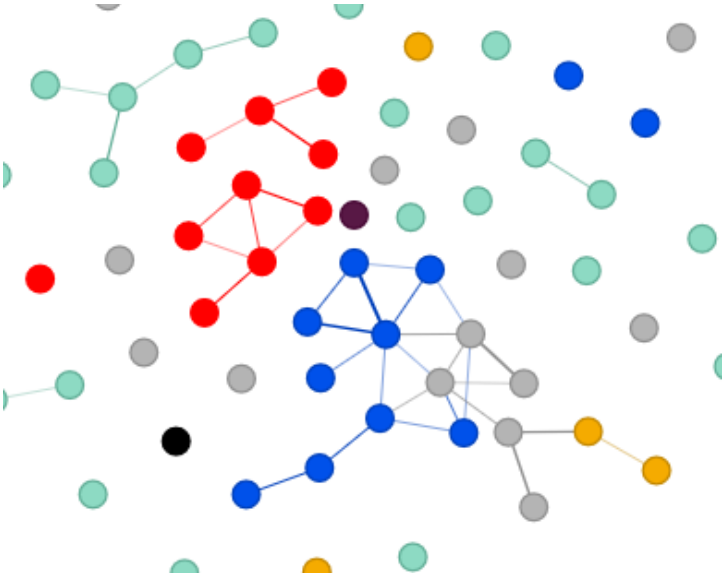

Рис. 1. Фрагмент графа связи сообществ

Как можно увидеть из рисунка 1, сообщества со схожими тематиками зачастую образуют связанные узлы в графе, что соответствует гипотезе о схожести их текстового контента.

Также на рисунке можно заметить наличие редких связей между сообществами с разными тематиками. Такие связи могут помочь исследователям выделять сообщества, обладающими смешанным контентом, выявление которых является ресурсоемкой задачей при изучении социальных сетей экспертными методами.

Заключение

В результате работы был разработан метод, позволяющий определить степень сходства текстовых коллекций на основе косинусного сходства их векторных представлений, полученных методом doc2vec. Приведен пример использования метода для определения сходства текстового контента сообществ социальной сети.

Исследование выполнено при финансовой поддержке ГЗ «Наука», в рамках проекта FSWW-2020-0014.

Список использованных источников

1. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // ICLR (Workshop Poster). – 2013.
2. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). – 2013. – P. 3111–3119.
3. Le Q., Mikolov T. Distributed Representations of Sentences and Documents // Proceedings of the 31st International Conference on Machine Learning. – 2014. – P. 1188-1196.