ОЦЕНКА СХОДСТВА НАБОРОВ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ НА БАЗЕ КОСИНУСНОГО СХОДСТВА И TF-IDF

А.О. Савельев, к.т.н., доцент ОИТ ИШИТР
С.А. Кузнецов
Томский политехнический университет
E-mail: ksa11@tpu.ru

Введение

Одним из вызовов автоматизированной обработки открытых интернет-данных является отсутствие в них чёткой структуры. Самым распространённым видом слабоструктурированного контента по-прежнему является текстовый несмотря на общее снижение его доли относительно аудио - и видео - контента. В результате чего, текстовый контент может являться основой инструмента для задач классификации или кластеризации, в том числе целей таргетированной рекламы, определения целевой аудитории, оценки сходства целевых аудиторий между собой и т.д.

Таргетинг (от англ. «target» — цель) — это рекламный механизм, позволяющий выделить целевую аудиторию и показать рекламу именно ей.

Автоматический анализ текстового контента подразумевает, в первую очередь, решение задачи его векторизации.

Целью данной работы являлась разработка алгоритма предварительной подготовки данных для дальнейшего анализа и решения задач классификации или кластеризации на базе текстового контента с применением метода TF-IDF.

В качестве инструмента разработки был использован язык программирования Python.

Описание алгоритма

Для семантического анализа текста часто используется метод TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса [2-4].

TF (или частота слова) – это отношение количества употреблений какого-либо слова к совокупному количеству слов документа. Следовательно, анализируется значимость слова t_i в одном отдельном документе.

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t - количество слов t в текстовом документе, а в знаменателе – общее количество слов.

IDF — это обратная частотность документов, с которой какое-либо слово упоминается в документах коллекции. Для любого уникального слова в пределах точной коллекции документов присутствует одно значение IDF.

$$idf(t,D) = \log \frac{|D|}{\left|\{d_i \in D | t \in d_i\}\right|},$$

где $|D|$ - количество документов в коллекции.

$\left|\{d_i \in D | t \in d_i\}\right|$ - количество документов из коллекции D, в которой встречается t (когда $n_t \neq 0$).

TF-IDF считается как произведение двух выражений:

$$tf - idf(t,d,D) = tf(t,d) * idf(t,D)$$

Больший вес в рамках одного документа в TF-IDF имеют слова с высокой частотой и с невысокой частотой использования в иных документах.

Таким образом было сделано предположение, что по рассчитанному вектору TF-IDF возможно оценить сходство документов.

Предварительно, экспертным путём сформирована выборка сообществ социальной сети, на основе постов которых сформированы текстовые документы для дальнейшего анализа.

Для решения задачи оценки сходства был разработан алгоритм, включающий следующие этапы:

1) Извлечение постов сообществ за один календарный год
2) Объединение постов за один календарный год в один для каждого сообщества
3) Предварительная обработка текста, включающая следующие подэтапы:

a. очистка текста от знаков пунктуации;

b. очистка текста от «стоп-слов» (предлоги, суффиксы, междометия, цифры, частицы и т.д.)

c. нормализация текста – приведение каждого слова в нормальную форму

4) Расчет показателей TF-IDF. TF рассчитывается для каждого документа корпуса. IDF рассчитывается на документах корпуса с добавлением дополнительных документов для максимального расширения словаря. В качестве дополнительных документов взята открытая база новостей lenta.ru.

5) Расчёт косинусного расстояния между документами с использованием метода cosine_similarity библиотеки Scikit-learn [5].

Пример фрагмента матрицы косинусных расстояний документов приведён на рис. 1.

	0	1	2	3	4
0	1.000000	0.070724	0.102171	0.088330	0.035729
1	0.070724	1.000000	0.380165	0.229498	0.098313
2	0.102171	0.380165	1.000000	0.432497	0.207085
3	0.088330	0.229498	0.432497	1.000000	0.110166
4	0.035729	0.098313	0.207085	0.110166	1.000000

Рис. 1. Фрагмент матрица косинусных расстояний документов

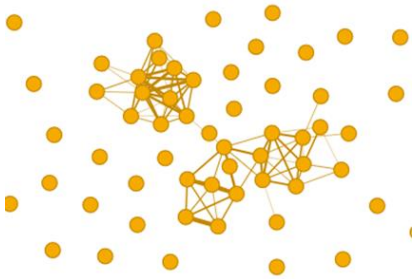

Рис. 2. Фрагмент графа связи документов

Для визуализации полученных результатов, построен граф связи документов (рис. 2).

На рис. 2 рёбра отражают наличие связи с весом выше порогового значения 0,7. Толщина ребра прямо пропорциональна значению связи документов.

Экспертный анализ полученных результатов подтвердил наличие значимой семантической связи между документами.

Заключение

В результате работы был разработан алгоритм, определяющий сходство между текстовыми документами на основе косинусного расстояния их векторизованных на базе TF-IDF-меры представлений.

Данный алгоритм может использоваться на этапе предварительной подготовки данных для решения задач классификации или кластеризации текстовой информации.

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

Список использованных источников

1. Kotinas I., Fakotakis N. Text Analysis for Decision Making under Adversarial Environments // Proceedings of the 10th Hellenic Conference on Artificial Intelligence. – 2018. – Article No.: 39. – P.1- 6
2. Miranda E., Aryuni M., Fernando Y., Kibtiah T. A Study of Radicalism Contents Detection in Twitter: Insights From Support Vector Machine Technique // 2020 International Conference on Information Management and Technology (ICIMTech). – 2020 – P. 549-554.
3. Machová K., Mach M., Demková G. Modelling of the Fake Posting Recognition in On-Line Media Using Machine Learning // SOFSEM 2020: Theory and Practice of Computer Science. - 2020 – P. 667-675.
4. Scikit-learn Machine Learning in Python [Электронный ресурс]. – Режим доступа URL: https://scikit-learn.org/stable/index.html (дата обращения 01.03.2021).