DOI: 10.51790/2712-9942-2023-4-1-05

АЛГОРИТМЫ ОБРАБОТКИ И ВЫЧИСЛЕНИЯ СХОДСТВА ТЕКСТОВЫХ ДАННЫХ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

Ш. И. Мутаиров[а], К. И. Бушмелева[б]
Сургутский государственный университет, г. Сургут, Российская Федерация
[а] ✍ *usermage@gmail.com,* [б] *bkiya@yandex.ru*

Аннотация: в статье описаны алгоритм вычисления сходства текстовых данных пользователей социальных сетей, алгоритм нормализации текста, а также алгоритм рекомендации проектов участникам образовательного процесса.

Ключевые слова: обработка естественного языка, косинусная мера.

Для цитирования: Мутаиров Ш. И., Бушмелева К. И. Алгоритмы обработки и вычисления сходства текстовых данных пользователей социальных сетей. *Успехи кибернетики*. 2023;4(1):33–38. DOI: 10.51790/2712-9942-2023-4-1-05.

Поступила в редакцию: 07.02.2023. В окончательном варианте: 21.02.2023.

TOOLS FOR SOCIAL MEDIA USER-GENERATED CONTENT SIMILARITY ASSESSMENT

Sh. I. Mutairov[a], K. I. Bushmeleva[b]
Surgut State University, Surgut, Russian Federation
[a] ✍ *usermage@gmail.com,* [b] *bkiya@yandex.ru*

Abstract: the paper describes the proposed algorithms for the similarity assessment in social media user-generated texts, text normalization, and project recommendation.

Keywords: natural language processing, cosine similarity.

Cite this article: Mutairov Sh. I., Bushmeleva K. I. Tools for Social Media User-Generated Content Similarity Assessment. *Russian Journal of Cybernetics*. 2023;4(1):33–38. DOI: 10.51790/2712-9942-2023-4-1-05.

Original article submitted: 07.02.2023. Revision submitted: 21.02.2023.

Введение

В последнее время в образовании все большее распространение получает так называемый проектный подход. Одной из задач проектного подхода является поиск участников, заинтересованных в реализации проекта и соответствующих предъявляемым требованиям. Рекомендательная система может быть использована для составления списка наиболее подходящих кандидатов для выполнения поставленной задачи, но для начала необходимо получить информацию о самих кандидатах [1]. В современную эпоху вездесущего социального взаимодействия через Интернет можно попытаться собрать необходимые данные об участниках образовательного процесса из социальных сетей. Пользователи социальных сетей склонны раскрывать много личной информации. Эта информация может быть представлена в различных структурированных и неструктурированных формах. Накопление, анализ и последующая выработка рекомендаций на основе этих проанализированных данных может стимулировать участников образовательного процесса к совместной продуктивной деятельности.

В качестве механизма взаимодействия участников образовательного процесса предполагается использовать проектируемую социальную сеть университета, способную формировать рекомендации для ее пользователей с целью увеличения количества и повышения качества их социальных и профессиональных связей [8]. Для формирования релевантных рекомендаций необходимо для каждого участника рассчитать коэффициенты попарного его сходства с другими участниками по определенным атрибутам, которые, в большинстве, могут быть записаны в текстовом формате [9]. Для вычисления сходства некоторых атрибутов были разработаны алгоритмы, которые описаны далее в статье.

Алгоритм вычисления сходства текстовых строк

Для вычисления сходства между текстовыми строками, например, тем выпускных квалификационных работ (ВКР) участников образовательного процесса используется следующий алгоритм (рис. 1)

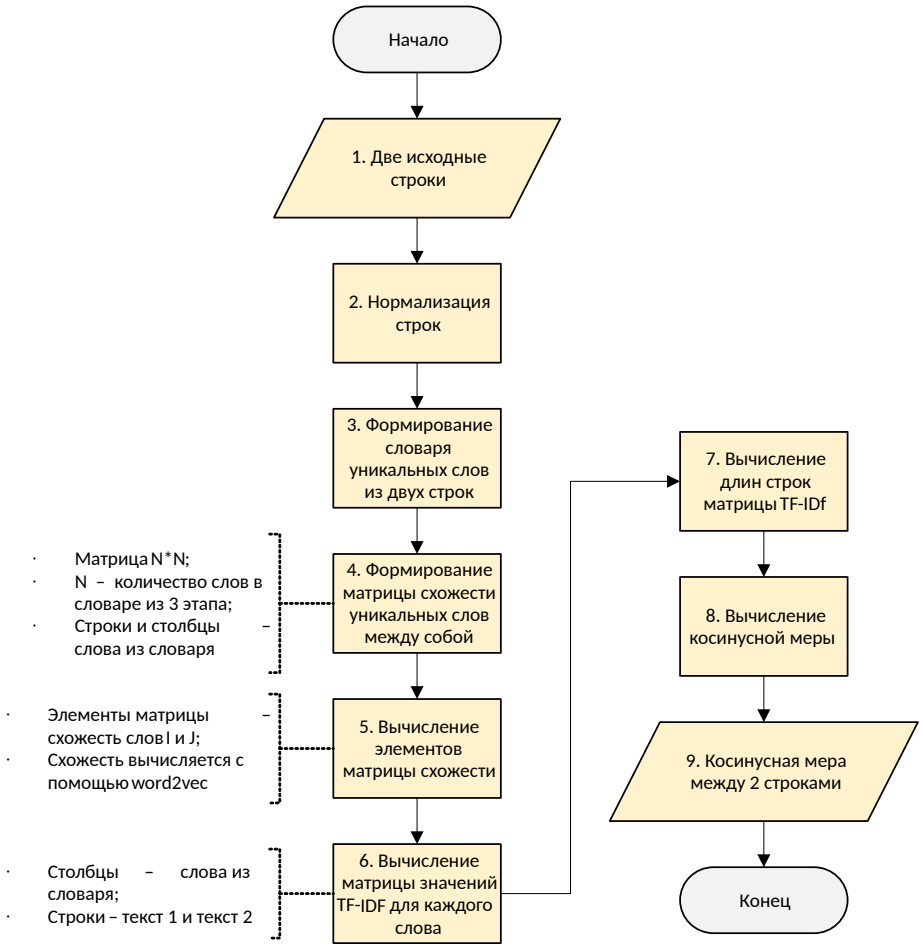

Рис. 1. *Алгоритм вычисления сходства строк*

На рисунке 2 изображен пример вычисления сходства двух строк. Для вычисления сходства между словами в матрице S (рис. 2) используется следующая система функций:

$$s_{ij} = \begin{cases} word2vec_{\text{сходство}}\left(\text{слово}_i, \text{слово}_j\right) & \text{сходство} \geq 0.6 \\ 0 & \text{сходство} < 0.6 \end{cases}. \quad (1)$$

Word2vec — это общее название для набора моделей на основе искусственных нейронных сетей, предназначенных для генерации векторных представлений слов на естественном языке. Он используется для семантического анализа естественного языка на основе дистрибутивной семантики, машинного обучения и векторного представления слов [3, 4]. Названное «word2vec» ПО было разработано группой исследователей в Google в 2013 году. Библиотека Gensim, написанная на Python, является одной из самых популярных библиотек для использования моделей word2vec.

Элементы в матрице значений TF-IDF (рис. 2) вычисляются по следующей формуле:

$$\frac{n_{t,d}}{n_d} * \log_{10}\left(\frac{N}{N_t}\right), \quad (2)$$

где $n_{t,d}$ — количество вхождений слова t в тексте d;

n_d — общее количество слов в тексте d;

N — общее количество текстов;

N_t — количество текстов, содержащих слово t.

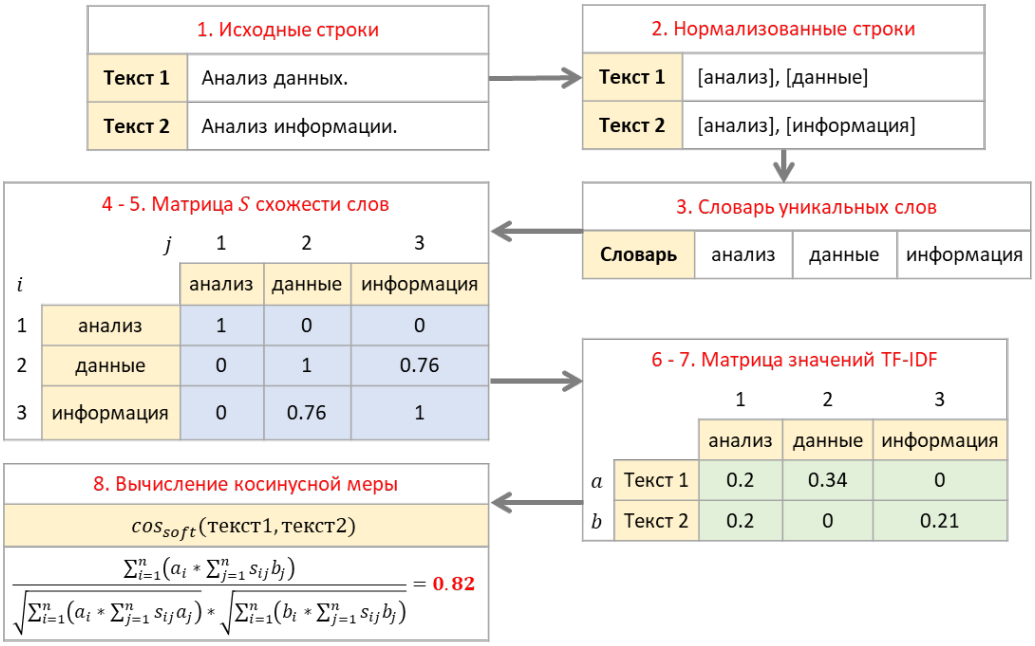

Рис. 2. *Пример вычисления сходства двух строк*

Алгоритм нормализации текста

В описанном выше алгоритме для повышения точности определения смыслового сходства двух строк выполняется их предварительная нормализация. Процесс нормализации позволяет убрать из исходного текста грамматическую информацию (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее), оставляя только смысловую составляющую [5, 6].

Алгоритм нормализации текста изображен на рисунке 3.

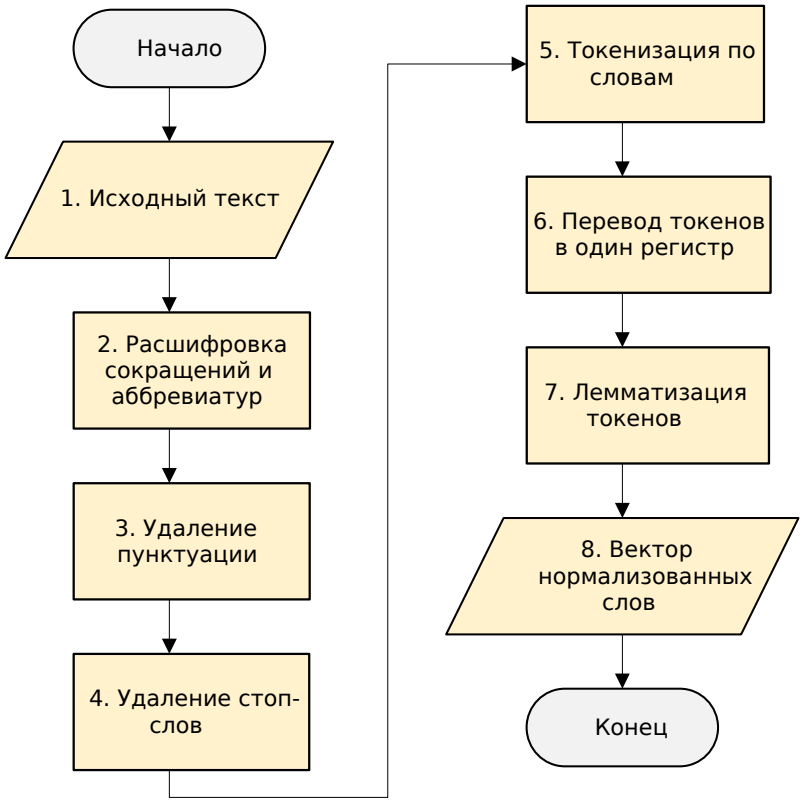

Рис. 3. *Алгоритм нормализации текста*

На шаге 2 расшифровка сокращений и аббревиатур проводится с помощью библиотеки Pymorphy2. Pymorphy2 – морфологический анализатор, написанный на языке Python (работает под 2.7 и 3.3+) [7]. Он умеет:

- приводить слово к нормальной (исходной) форме (например, «люди → человек» или «гулял → гулять»);
- ставить слово в нужную форму. Например, ставить слово во множественное число, менять падеж слова и т.д.;
- возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.).

При работе используется словарь OpenCorpora; для незнакомых слов строятся гипотезы. Библиотека работает очень быстро: в настоящее время скорость работы составляет от нескольких тыс. слов/сек до свыше 100 тыс. слов/сек (в зависимости от выполняемой операции, интерпретатора и установленных пакетов); потребление памяти составляет 10 … 20 Мб; полностью поддерживается буква ё.

На шаге 4 с помощью свободной библиотеки NLTK удаляются стоп-слова. NLTK предназначена для символьной и статистической обработки естественного языка и написана на языке программирования Python. Отдельно искать словари стоп-слов не требуется, поскольку в NLTK они уже предустановлены, причем для многих языков мира, включая и русский.

Стоп-слова — это слова, которые удаляются из текста до или после его обработки, как правило, это союзы, междометия, артикли и т.д., не несущие семантической нагрузки. При применении машинного обучения в обработке текстов необходимо удалять нерелевантные слова, поскольку такие слова создают много шума.

На шаге 5 токенизация по словам проводится с помощью Pymorphy2. Токенизация (иногда называемая сегментацией) по словам — это процесс разбиения предложения на составляющие его слова. В языках (например, в английском), в которых используется та или иная версия латинского алфавита, пробел является эффективным разделителем слов. Однако если производить разбиение текста только по пробелу, то может быть утрачен смысл составных существительных или словосочетаний, поэтому токенизацию необходимо выполнять с учетом контекста — Pymorphy2 это умеет.

На шаге 7 лемматизация токенов проводится тоже с помощью Pymorphy2. Цель лемматизации — привести все встречающиеся словоформы к одной, нормальной (исходной) словарной форме для упрощения процесса программного анализа текста.

В таблице 1 продемонстрирован пример работы алгоритма нормализации текста.

Таблица 1

Пример работы алгоритма нормализации текста

№	**Операция**	**Результат**
1	Исходный текст	Александр учится на направлении ИВТ.
2	Расшифровка сокращений и аббревиатур	Александр учится на направлении информатика и вычислительная техника.
3	Удаление пунктуации	Александр учится на направлении информатика и вычислительная техника
4	Удаление стоп-слов	Александр учится на направлении информатика и вычислительная техника
5	Токенизация по словам	[Александр] [учится] [направлении] [информатика] [вычислительная техника]
6	Перевод токенов в один регистр	[александр] [учится] [направлении] [информатика] [вычислительная техника]
7	Лемматизация токенов	[александр] [учиться] [направление] [информатика] [вычислительный] [техника]
8	Вектор нормализованных слов	<[александр], [учиться], [направление], [информатика], [вычислительный], [техника]>

Алгоритм рекомендации проектов пользователю

Суть алгоритма заключается в следующем:

1. Берется вектор проектов $\overrightarrow{P}$ и вектор интересов $\overrightarrow{Y}$ пользователя.
2. Составляется матрица A, где строки — это вектор проектов, а столбцы — вектор интересов.
3. Вычисляются элементы матрицы A по следующей формуле:

$$a_{ij}=\frac{n_{j,i}}{n_i}*\log_{10}\left(\frac{N}{N_j}\right)*V, \tag{3}$$

где $n_{j,i}$ — количество вхождений интереса j в тексте (название, описание) проекта i;
n_i — общее количество слов в описании проекта i;
N — общее количество проектов;
N_j — количество проектов, содержащих в описании интерес j;
V — вес интереса {высокий — 3, средний — 2, низкий — 1}.

4. Вычисляется вектор $\overrightarrow{H}$ средних значений a_{ij}:

$$h_i=\sum_j a_{ij.} \tag{4}$$

5. Вычисляется вектор $\overrightarrow{S}$ сходства проектов и темы ВКР:

$$s_i=\text{сходство }(y_i,\text{тема ВКР}).$$

Для этого используется ранее описанный алгоритм вычисления сходства текстов.

6. Вычисляется вектор коэффициентов рекомендаций $\overrightarrow{R}$ по формуле:

$$r_i=(h_i+s_i)/2. \tag{5}$$

7. Сортируется (по убыванию) вектор проектов по их коэффициентам рекомендации, следует понимать, что i-ый проект имеет i-ый коэффициент рекомендации.
8. Пользователю отображается отсортированный вектор проектов.

Выполнение описанных выше шагов проиллюстрировано на рисунке 4.

Пример матрицы модифицированных значений TF-IDF интересов

i / j		1	2	3	4			
		Интерес 1	Интерес 2	Интерес 3	Интерес 4	Среднее коэфф-ов	Схожесть с темой ВКР	Коэфф. рекомендации
1	Проект 1	0.6	0.3	0.6	0.2	0.4	0.7	0.57
2	Проект 2	1.2	0.7	1.3	0.7	1.0	0.42	0.70
3	Проект 3	0.4	1.0	0.9	0.2	0.6	0.65	0.64

N_t	4	2	3	2
N	100			

Итоговый список рекомендаций

Место	Проект
1	Проект 2
2	Проект 3
3	Проект 1

Рис. 4. *Алгоритм рекомендации проектов пользователю*

В зеленых ячейках таблицы (рис. 4) содержатся коэффициенты TF-IDF, вычисленные по формуле 3, в синих — средние коэффициентов TF-IDF (по строке), в оранжевых — степени схожести названий

проектов с названием темы ВКР, в розовых — коэффициенты рекомендации проектов, вычисленные по формуле 5. В итоговой таблице представлен отсортированный по рекомендуемости список проектов.

Заключение

В результате проведенных исследований разработаны алгоритм вычисления сходства текстовых строк, алгоритм нормализации текста, алгоритм рекомендации проектов пользователю. Данные алгоритмы являются частью метода автоматизированного вычисления интегрального показателя сходства участников образовательного процесса по различным атрибутам [8–9].

ЛИТЕРАТУРА

1. Анатомия рекомендательных систем. Часть первая. *Habr*. Режим доступа: https://habr.com/ru/company/lanit/blog/420499/.
2. Sidorov G. et al. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas*. 2014;18(3):491–504.
3. *Векторная модель*. Режим доступа: https://ru.wikipedia.org/wiki/Векторная_модель#«Мягкая»_косинусная_мера.
4. *Word2vec*. Режим доступа: https://ru.wikipedia.org/wiki/Word2vec.
5. Основы Natural Language Processing для текста. *Habr*. Режим доступа: https://habr.com/ru/company/Voximplant/blog/446738/.
6. Математические методы анализа текстов Семинар 1. *Machinelearning*. Режим доступа: http://www.machinelearning.ru/wiki/images/5/53/Mel_lain_msu_nlp_sem_1.pdf.
7. Морфологический анализатор pymorphy2. *Pymorphy2*. Режим доступа: https://pymorphy2.readthedocs.io/en/0.2/user/index.html.
8. Mutairov Sh. I., Bushmeleva K. I. Method and Algorithms for Organizing Joint Activities of Participants in the Educational Process. *Information Innovative Technologies*. Prague: International scientific-practical conference; 2022.
9. Мутаиров Ш. И., Бушмелева К. И. Алгоритмы вычисления сходства пользователей социальной сети. *Инновационные, информационные и коммуникационные технологии*: сборник трудов XIX Международной научно-практической конференции / под. ред. С. У. Увайсова. Москва: Ассоциация выпускников и сотрудников ВВИА им. проф. Жуковского; 2022. С. 388.