Вычислительные технологии, 2020, том 25, № 5, с. 107–123. © ФИЦ ИВТ, 2020
Computational Technologies, 2020, vol. 25, no. 5, pp. 107–123. © FRC ICT, 2020
ISSN 1560-7534
eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2020.25.5.009

# Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов

Ф. В. Краснов*, И. С. Смазневич
NAUMEN R&D, Екатеринбург, Россия
*Контактный автор: Краснов Федор Владимирович, e-mail: fkrasnov@naumen.ru

*Поступила 26 августа 2020 г., доработана 21 сентября 2020 г., принята в печать 25 сентября 2020 г.*

С развитием все более сложных методов автоматического анализа текста повышается важность задачи объяснения пользователю, почему прикладная интеллектуальная информационная система выделяет некоторые тексты как схожие по смыслу. В работе рассмотрены ограничения, которые такая постановка накладывает на используемые интеллектуальные алгоритмы.

Проведенный авторами эксперимент показал, что абсолютное значение схожести документов не универсально по отношению к интеллектуальному алгоритму, поэтому оптимальную пороговую величину схожести необходимо устанавливать отдельно для каждой решаемой задачи.

Полученные результаты могут быть использованы при оценке применимости различных методов установления смысловой схожести между документами в прикладных информационных системах, а также при выборе оптимальных параметров модели с учетом требований объяснимости решения.

*Ключевые слова*: explainable artificial intelligence, XAI, функция ранжирования, схожесть документов.

*Цитирование*: Краснов Ф.В., Смазневич И.С. Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов. Вычислительные технологии. 2020; 25(5):107–123. DOI:10.25743/ICT.2020.25.5.009.

## Введение

Интеллектуальные информационные системы (ИИС) развиваются с 70-х гг. XX в. Под интеллектуальностью в рамках настоящей работы следует понимать согласно Д. А. Поспелову [1] “сознание” вычислительной машины или комплекса, т. е. способность имитировать аналитическую деятельность человека. Разработки в области ИИС исторически велись в двух основных направлениях:

- создание ИИС путем моделирования их биологического прототипа — человеческого мозга;
- создание ИИС, обеспечивающих решение сложных математических и логических задач, позволяющих автоматизировать отдельные интеллектуальные действия человека.

Эти две цели как бы определяют программу максимум и программу минимум, между которыми и лежит область сегодняшних исследований и разработок ИИС. Далее будет рассмотрено второе направление, которое развивается в прикладных интеллектуальных информационных системах (ПИИС).

ПИИС могут использоваться в различных условиях: для персональных целей (например, при поиске информации в интернете — в Википедии, словарях) и при решении корпоративных задач сотрудниками организации. При корпоративном использовании ПИИС подразумевается, что человек на рабочем месте участвует в бизнес-процессе компании и взаимодействует с системой для выполнения бизнес-операций. К таким современным ПИИС можно отнести доменные семантические порталы научно-технической информации [2], систему автоматизации рецептурного производства [3], систему поддержки принятия решений при управлении программными проектами на основе нечеткой онтологии [4].

Перечисленные ПИИС объединяет общий методический подход, который подробно рассмотрен в исследовании [5]: на определенном шаге бизнес-процесса в случае принятия сотрудником решения учитывается результат работы интеллектуального алгоритма (ИА) — составляющей прикладной интеллектуальной информационной системы. В этом случае результат работы носит локальный рекомендательный характер. Локальность означает, что результат интеллектуального алгоритма зависит от времени $t_0$ и от роли сотрудника $r_0$ в ПИИС. Другими словами, для другой роли сотрудника $r_1$ и/или в другое время $t_1$ на этом же шаге бизнес-процесса интеллектуальный алгоритм может выдать в качестве рекомендации другой результат. Рекомендательный характер ПИИС является их основным отличием от систем, автоматизирующих бизнес-процессы, таких как системы электронного документооборота, системы управления проектами и системы полнотекстового поиска.

**Пример 1.** У компании есть $N = 100\,000$ нормативно-технических документов. Методист $r_0$ собирается разработать новый нормативно-технический документ $D_0$ с помощью ПИИС “АИС НТД”. Методист создал проект (черновик) документа $D_0$ и сохранил его в ПИИС. На этом шаге бизнес-процесса $t_0$ ПИИС “АИС НТД” рекомендует обратить внимание на схожие документы (как частично совпадающие по тексту, так и близкие по смыслу) и предлагает их список, отсортированный по степени схожести $D_n$. Методист анализирует предложенные документы и делает вывод (принимает решение) о том, как ему следует разрабатывать документ $D_0$, чтобы использовать уже существующую нормативную базу без противоречий и дублирования.

Разберем подробнее поведение ПИИС “АИС НТД” в примере 1. Интеллектуальный алгоритм в данном случае определяет схожесть документов коллекции (нормативно-методической базы) с документом $D_0$, имитируя последовательные действия человека: выбрать документ $D_i$, прочитать документ $D_i$, понять смысл $\mathbf{T}_i$ документа $D_i$, сравнить со смыслом $D_1$, определить скалярную меру схожести документов $s_{0i} = S(\mathbf{T}_0, \mathbf{T}_i)$. И так алгоритм действует со всеми документами $D_n$.

Для интеллектуального алгоритма, в отличие от человека, не составляет проблемы вычислить векторы $\mathbf{T}_i$ для всех документов $D_i$ и запомнить их как пространство смыслов $H_D^T$. Таким образом, под смысловой схожестью документов можно понимать величину, обратно зависимую от расстояния $d_{ij}$ между их векторными представлениями в пространстве $H_D^T$. Чем это расстояние $d$ меньше, тем более схожи документы по смыслу: $s_{ij} = 1 - d_{ij}$.

Важно отметить, что пространство смыслов $H_D^T$ может быть недоступно пользователю для понимания, однако ему понятна конкретная рекомендация интеллектуального алгоритма — список схожих по смыслу документов, отсортированный по степени схожести с данным документом. При этом возможность объяснения неочевидного результата, полученного алгоритмом рекомендательной системы, зачастую является одним

из критериев, по которым компания — заказчик прикладной информационной системы — делает выбор в пользу приобретения того или иного технологического решения (ПИИС), а также более или менее активно тестирует систему в период опытной эксплуатации.

Достаточно сложно найти емкое определение для "необъяснимости алгоритма". Один из побочных эффектов теорем Гёделя 1931 г. состоит в том, что даже если у утверждения есть доказательство, путь к нему может быть сколь угодно длинным. Таким образом, необъяснимость можно характеризовать как порог по длительности времени, требуемого для объяснения. Этот порог варьируется в рассматриваемом нами случае от интуитивно понятного пользовательского интерфейса ПИИС, не требующего дополнительного обучения, до необходимости длительного корпоративного обучения пользователей при внедрении ПИИС. Теоретические основы "необъяснимости алгоритма" лежат за рамками настоящего исследования, но с точки зрения пользы от ПИИС нам не так важно пытаться объяснить, каким образом результат был получен на самом деле, как использовать для получения результата операции, уже известные пользователю.

Целью данного исследования является поиск интеллектуального алгоритма, оптимального с точки зрения возможностей по объяснимости сделанных им рекомендаций. Статья включает помимо вводной части обзор существующих методов решения задач по работе с текстовыми документами, в том числе задачи кластеризации, с учетом объяснимости этих методов; описание проведенного эксперимента по определению схожих документов с помощью различных алгоритмов; заключительный раздел с выводами о применимости различных методов в ПИИС, где требуется объяснимость предложенных пользователю рекомендаций.

## 1. Методика

Продемонстрированный в примере 1 подход к решению прикладных информационных задач, известный как "человек в цикле" (human in the loop) [6–8], сочетает преимущества человеческого разума и машинного интеллекта. Машины хороши для принятия "умных" решений на основе обширных наборов данных, тогда как люди гораздо лучше принимают решения при меньшем количестве информации. Например, умеют определять отдельные значимые фрагменты документа (сущности), не вчитываясь в текст ("это шапка документа", "это термины", "это код документа"), или способны с первого взгляда классифицировать документ как протокол совещания.

Чтобы сымитировать работу человека, интеллектуальный алгоритм должен уметь похожим образом выделять необходимые данные из текстового слоя документа. То, как будет реализован этот шаг — выделение сущностей или разбиение текста на значимые фрагменты (структурирование документа), — вносит значительный вклад в объяснимость конечного результата работы алгоритма.

Поскольку принцип "человек в цикле" на сегодняшний день является наилучшей практикой применения искусственного интеллекта (реализованной, например, в рекомендательных ИС) и промежуточным шагом на пути к "сильному искуственному интеллекту", то важно отметить, что данная схема работы может быть оправданна и эффективна только при достаточной степени "прозрачности" выводов алгоритма, способной обеспечить необходимый уровень доверия пользователя к рекомендациям системы.

Существующие подходы к выделению информации из текста основываются на анализе естественного языка [9], построении правил [10] и выделении семантических при-

знаков [11]. Результаты выделения информации из текста могут быть использованы по-разному: для формирования запроса на поиск во внешних источниках [12], для сравнения двух документов по выделенной информации [13] или для представления выделенной информации пользователю [14].

В ПИИС за выделение информации из текста отвечает интеллектуальный алгоритм, важной характеристикой которого является самообъяснимость — свойство алгоритма выдавать результат, происхождение и логика получения которого понятны человеку. Такая особенность алгоритма позволяет избежать известной проблемы "черного ящика" в машинном обучении [15], из-за которой даже разработчики не всегда могут объяснить, почему эстиматор пришел к определенному решению. Растущая вычислительная способность машин и увеличивающаяся сложность алгоритмов привели к необходимости введения понятия "объяснимого искусственного интеллекта" (Explainable Artificial Intelligence, XAI). Для прицельной разработки решений такого типа требуется ввести критерии отнесения интеллектуальных систем к "объяснимым".

В работе [16] понятие объяснимости формализуется через оценку этой характеристики алгоритма по двум критериям: интерпретируемости и полноте описания.

Задача объяснения решений ИА может быть сведена к интерпретации [17], т. е. представлению результата в понятных пользователю терминах. Когда объем информации, подлежащей объяснению, получается слишком велик, необходимо представить ее в удобном для понимания виде, например сделать проекцию векторов большой размерности на двухмерное пространство [18]. Однако интерпретация зависит в том числе и от самого пользователя: термины, понятные одному человеку, другому могут показаться трудными для восприятия. Поэтому важно понимать, что интерпретируемость не должна достигаться за счет чрезмерного упрощения. Стремясь завоевать доверие пользователей, следует избегать введения их в заблуждение (излишне редуцируя сложные для понимания данные).

С другой стороны, в качестве объяснения результата может выступать само описание характера и логики работы ИА. По сути, алгоритм перебирает параметры модели, чтобы удовлетворить математически заданной целевой функции, выбранной разработчиками системы (например, минимизировать информационную энтропию языковой модели).

Попытка объяснения пути поиска тоже может служить обоснованием найденного оптимума, формируя у пользователей уверенность в том, что ИА обязательно переберет все варианты. Представление пути поиска является достаточно наглядным в случае градиентного спуска [19], однако результаты генетических методов [20] и "оптимизации роем частиц" [21] объяснить гораздо сложнее, поэтому такие методы вызывают меньше рационального доверия у пользователей.

Таким образом, при выборе наилучшего метода для решения оптимизационной задачи объяснимость решения становится одним из критериев наряду с эффективностью интеллектуального алгоритма. При этом необходимо рассматривать такое объяснение, в котором достигнут компромисс между простотой интерпретации и полнотой описания.

Для работы с текстами направление "объяснимый искусственный интеллект" представлено методикой LIME (Local Interpretable Model-Agnostic Explanations) [22]. В задаче классификации она позволяет объяснять предсказания произвольных классификаторов, в том числе текстовых, поскольку демонстрирует точное соответствие между коэффициентами модели и текстовыми элементами.

Для решения задачи кластеризации рассмотрение пространства смыслов приводит к следующим основным методам: LSI (Latent Semantic Indexing) [23], LDA (Latent Dirichlet Allocation) [24], NMF (Non-negative Matrix Factorization) [25], ARTM (Additive Regularization of Topic Models) [26].

Необходимо отметить, что за рамками настоящего исследования остается класс ресурсоемких методов WMD (Word Mover's Distance) [27], в алгоритме которых можно использовать эмбедингги для лемм: статические (Word2Vec, GloVe, FastText) и динамические (Elmo, BERT).

Качество полученного решения в задаче кластеризации зависит не только от выбранного метода, но и от заданной метрики ($S(\circ,\circ)$), определяющей степень близости между объектами пространства. Исчерпывающее исследование метрик для сравнения предложений, слов и текстов выполнено в работе [28], однако здесь не затронут вопрос объяснимости метрик. При этом понимание того, как работает метрика пространства $H_D^T$, в котором сравниваются тексты, является обязательным условием для ответа на вопрос, почему ИА считает два текста схожими по смыслу. Наиболее часто применяются косинусная и евклидова метрики пространства:

$$S_E(\mathbf{T}_i,\mathbf{T}_j)=1-\sqrt{\frac{\sum\limits_{l=0}^{k}(t_{il}-t_{jl})^2}{\sum\limits_{l=0}^{k}t_{il}^2\sum\limits_{l=0}^{k}t_{jl}^2}},\qquad S_{\cos}(\mathbf{T}_i,\mathbf{T}_j)=\frac{\sum\limits_{l=0}^{k}(t_{il}*t_{jl})}{\sqrt{\sum\limits_{l=0}^{k}t_{il}^2\sum\limits_{l=0}^{k}t_{jl}^2}},$$

где $S_E(\mathbf{T}_i,\mathbf{T}_j)$ и $S_{\cos}(\mathbf{T}_i,\mathbf{T}_j)$ — евклидова и косинусная меры схожести между векторами $\mathbf{T}_i$ и $\mathbf{T}_j$ в пространстве смыслов $T$ размерности $k$.

Косинусная и евклидова меры схожести линейно связаны между собой. Обе эти метрики достаточно просто объясняются геометрически — на примере двухмерного пространства. Однако обычно на практике размерность пространства составляет несколько сотен измерений. В таких случаях обойтись без геометрической интерпретации при объяснении позволяет, например, частотная метрика Jaccard Score [29], определяющая схожесть сравниваемых текстов по количеству совпадающих в них слов. Это же объяснение подходит и при анализе схожести с помощью ранжирующих функций на основе TF-IDF [30] и более современных — на основе BM25 [31], которые позволяют за счет нормирования векторов выделить в документах слова, наиболее значимые для сравнения. Существует множество вариаций ранжирующих функций, например в работе [32] рассмотрены пять вариантов реализации ранжирующих функций (BM25, BM25+, BM25T, BM25-adpt, BM25L) и показано, что между ними нет системной разницы в результатах.

## 2. Эксперимент

Целью эксперимента было определение оптимальных методов и параметров алгоритма для решения задачи по нахождению схожих документов в коллекции, т. е. таких характеристик ИА, при которых решение этой задачи становится наиболее объяснимым. В качестве экспериментального корпуса текстов выбран корпус "Тайга" [33] из 7696 документов, словарь которых содержит 15 095 слов. В среднем один документ состоит из 211 слов со среднеквадратическим отклонением 103.5 слова. Для приведения к нормальной форме слов использовалась библиотека PyMorphy2 [34].

Исследованы матрицы схожести текстов, полученные с помощью ранжирования TF-IDF и вычисления тематических компонентов поочередно с помощью следующих четырех алгоритмов: LSI, LDA, NMF и ARTM. В качестве метрик векторного пространства использованы как евклидово расстояние, так и косинусная мера. В первую очередь производилось ранжирование массива документов с помощью TF-IDF, а затем использовалось уменьшение размерности векторной модели (матрицы) до заданного значения с помощью LSI, LDA, NMF, ARTM. На рис. 1 и 2 изображены распределения степени схожести между документами для различных методов обработки текста.

С точки зрения объяснимости алгоритма преимущество TF-IDF состоит в идентификации наборов слов, которые являются дискриминационными для документов в коллекции. Однако метод дает сравнительно небольшое уменьшение размерности векторов, описывающих документы (размерность уменьшается только за счет фильтрации словаря), и мало что сообщает о статистической структуре документа. Чтобы устранить эти недостатки, были предложены методы уменьшения размерности, в частности скрытое семантическое индексирование (LSI) [23].

LSI использует разложение матрицы "документ-словарь" по сингулярному значению (Singular Value Decomposition, SVD) для идентификации линейного подпространства в пространстве функций TF-IDF, которое фиксирует бóльшую часть дисперсии в коллекции:

$$X = U\Sigma V^{-1},$$

где $X$ — исходная матрица коллекции "документ-словарь"; $\Sigma$ — матрица с сингулярными числами на главной диагонали и нулевыми остальными элементами; $U$ и $V$ — унитарные матрицы, состоящие из левых и правых сингулярных векторов соответственно.

Этот подход позволяет добиться значительного сжатия разреженных векторных данных в больших коллекциях за счет замены исходных функций TF-IDF на их линейные комбинации в модели LSI. Кроме того, авторы работы [23] утверждают, что метод LSI может выделять некоторые аспекты основных языковых понятий, таких как синонимия и полисемия.

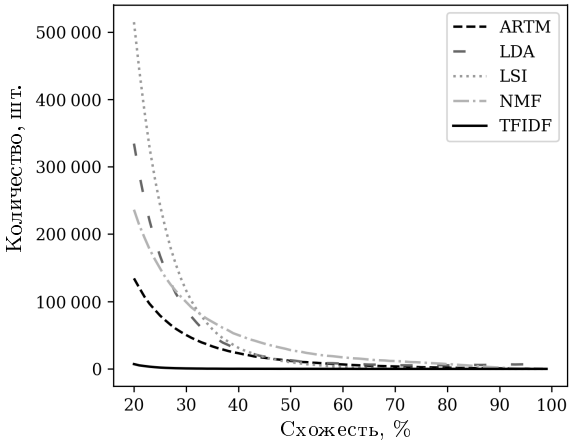


Рис. 1. Распределение количества документов разной степени схожести для различных методов с использованием косинусной метрики
Fig. 1. Distribution of document quantities that have different degrees of similarity for various methods using the cosine metric

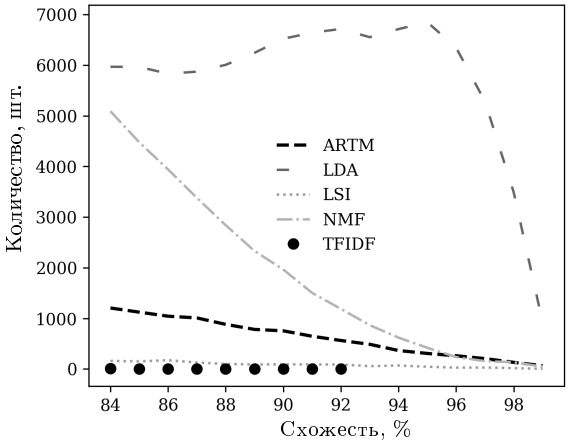


Рис. 2. Распределение количества документов со степенью схожести от 84 % для различных методов с использованием косинусной метрики
Fig. 2. Distribution of document quantities with degree of similarity greater 84 % for various methods using the cosine metric

В рамках настоящего исследования в качестве реализации алгоритма LSI выбран оптимизированный под параллельные вычисления вариант SVD-преобразования с использованием рандомизации (RSVD) [35]. Для этого алгоритма планку точности задает величина объясняемой дисперсии, что в данном случае означает долю словаря, учтенную при формировании тематик корпуса. Зависимость объясняемой дисперсии и остаточной вариативности матрицы $\Sigma$ от размерности $k$ для компактного преобразования представлена на рис. 3.

Рекомендуемое в научной литературе [23, 36] значение для $k \in [50, \ldots, 300]$ при размерности словаря $m = 1.5 \cdot 10^4$ слов оставляет около 80 % дисперсии без объяснения (эти слова не учитываются в тематиках). Поэтому в проведенном эксперименте для методов LSI (SVD) и NMF использовано значение $k = 3000$, при котором необъясненными остаются лишь 20 % дисперсии.

При переходе от ранжирующих функций к методам тематического моделирования объяснение решения (вывода о схожести документов) перестает быть очевидным, поэтому в качестве аналога объяснимости алгоритма теперь может выступать когерентность тематик. Эта величина является формальной метрикой их качества и может служить критерием для оценки алгоритма с точки зрения его объяснимости.

Существует несколько методик определения когерентности [37, 38]. Общий подход к ее вычислению состоит в том, чтобы рассчитывать частоту парного появления слов в документах и нормировать их частотой появления в корпусе каждого из слов по отдельности.

Преимущества различных подходов к вычислению когерентности в научной литературе не обоснованы, поэтому в настоящем исследовании использовалась следующая формула для вычисления когерентности тематики, достаточно удобная с точки зрения объяснимости (поскольку не ухудшает общую объяснимость алгоритма):

$$C_t = \frac{2}{N_t\,(N_t - 1)} \sum_{w_i \in t} \sum_{w_j \in t} \log \left[ \frac{p(w_i, w_j)}{p(w_i) p(w_j)} \right], \quad i \neq j, \tag{1}$$

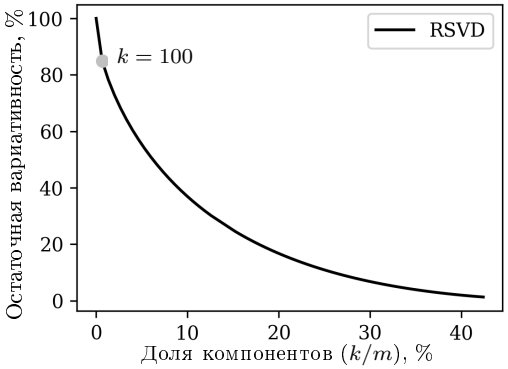


Рис. 3. Зависимость остаточной вариативности матрицы $U$ от доли компонентов $k$ в словаре $m$

Fig. 3. Dependence for the residual variance of the matrix $U$ on the percentage of components $k$ in the dictionary $m$

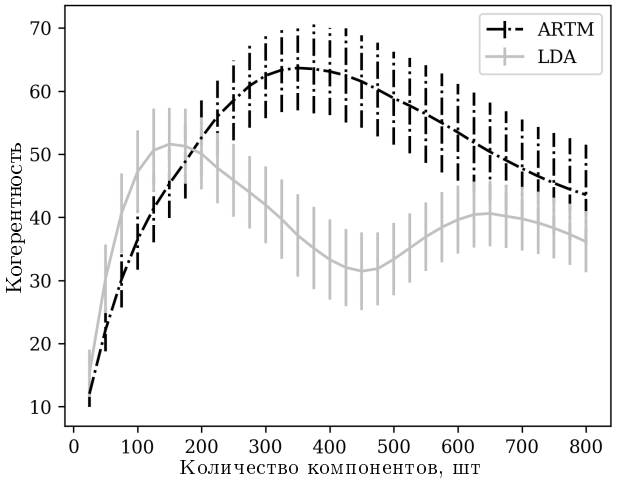


Рис. 4. Зависимость средней когерентности от количества тематических компонентов $k$ для алгоритмов LDA и ARTM

Fig. 4. Dependence of the average coherence on the number of topics (components) $k$ for the LDA and ARTM algorithms

Количество схожих пар документов при использовании различных методов и метрик схожести
The number of similar pairs of document when using various similarity methods and metrics

| Схожесть, | Евклидово расстояние | | | | Косинусная мера | | | |
|---|---|---|---|---|---|---|---|---|
| % | ARTM | LDA | NMF | LSI | ARTM | LDA | NMF | LSI |
| 99 | 0 | 1 | 0 | 0 | 232 | 3575 | 45 | 0 |
| 98 | 0 | 6 | 2 | 0 | 137 | 7560 | 133 | 0 |
| 97 | 2 | 29 | 3 | 0 | 200 | 9213 | 155 | 0 |
| 96 | 6 | 81 | 10 | 0 | 274 | 9049 | 235 | 1 |
| 95 | 8 | 183 | 12 | 0 | 326 | 9116 | 418 | 2 |
| 94 | 15 | 494 | 18 | 0 | 438 | 9045 | 619 | 7 |
| 93 | 14 | 1054 | 37 | 0 | 520 | 9222 | 865 | 5 |
| 92 | 37 | 1898 | 90 | 0 | 588 | 9038 | 1186 | 9 |
| 91 | 33 | 3153 | 142 | 0 | 715 | 8518 | 1503 | 11 |
| 90 | 50 | 4823 | 288 | 0 | 806 | 8261 | 1964 | 11 |

где $p(w_i)$ — частота употребления слова $w_i$ в корпусе текстовых документов по отношению к количеству слов в нем; $p(w_i, w_j)$ — относительная совстречаемость слов $w_i$ и $w_j$, вычисляемая как сумма отношений числа одновременных вхождений в документы слов $w_i$ и $w_j$ к количеству слов в каждом документе, где они встречались. При вычислении когерентности не учитывается совстречаемость слов с самими собой: $i \neq j$.

В выражении (1) когерентность тематики $t$ вычисляется по $N_t$ наиболее значимым для нее словам. Величина $N_t$ зависит от свойств конкретной тематики, поскольку эти слова отбираются по пороговому значению их веса в векторе тематики (составляющему в данном эксперименте 1 %). Умножение суммы на $\dfrac{2}{N_t(N_t-1)}$ делается по соображениям нормировки, так как перебирается размещение $C_2^N$ вариантов, т. е. сопоставляется “каждое слово с каждым”.

Распределение когерентности по тематикам носит нормальный характер согласно тесту Шапиро — Уилка. Поэтому для оценки качества построенной тематической модели может быть использована средняя когерентность по всем найденным тематикам для данной модели. С помощью максимизации средней когерентности тематик было определено оптимальное количество тематических компонентов для генеративных методов LDA и ARTM. Графики зависимости средней когерентности от числа тематик для этих алгоритмов приведены на рис. 4. Из графиков видно, что максимальная средняя когерентность в проведенном эксперименте достигалась при $k = 150$ компонентам для LDA и $k = 375$ компонентам для ARTM.

Важно отметить, что для всех интеллектуальных алгоритмов оптимальное количество компонентов разное. В дальнейшем для каждого алгоритма были выбраны оптимальные параметры, при которых рассчитывались по две матрицы схожести: по косинусной мере и евклидовому расстоянию. Результаты исследования приведены в таблице, из которой видно, что степень схожести не является универсальным показателем для всех методов тематического моделирования: тем парам документов, которые схожи на 99 % по LDA, не обнаруживается соответствующих схожих пар по LSI. Этот факт важен при решении различных задач в ПИИС, когда используются пороговые значения схожести, например, производится тематическая кластеризация коллекции по матрице схожести для дальнейшей маршрутизации документов или определяются похожие документы для рекомендации их пользователю.

**Пример 2.** Примером высокой степени схожести по TF-IDF являются следующие два документа.

- Программа AlphaGo выиграла последнюю игру против Ли Седоля — одного из сильнейших в мире игроков в го. Таким образом, компьютер выиграл четыре из пяти игр. Ли Седоль играл черными камнями, AlphaGo — белыми. За игрой можно было следить в прямой трансляции.
- Программа AlphaGo выиграла третью из пяти игр против Ли Седоля — одного из сильнейших в мире игроков в го. Таким образом, компьютер уже выиграл вне зависимости от исхода оставшихся двух партий. Ли Седоль играл черными камнями, AlphaGo — белыми. За игрой можно было следить в прямой трансляции.

Для данных текстов схожесть по алгоритму TF-IDF составила 84 %. При взгляде на эти тексты нетрудно заметить множество совпадающих слов, и это наблюдение иллюстрирует, что при достаточно высокой степени схожести по TF-IDF у человека не возникает сомнения относительно правильности сделанного алгоритмом вывода.

С помощью порогового значения были определены слова, которые внесли наибольший вклад в определение схожести между этими двумя документами (в порядке убывания важности для схожести): “седоль”, “alphago”, “выиграть”, “игра”, “го”, “сильнейший”, “ли”.

Как видно, именно эти слова определяют бóльшую часть смысла обоих документов, т. е. можно сказать, что они в совокупности отвечают на вопрос: о чем этот текст? Причем первые два слова — название программы и фамилия игрока — являются наиболее характерными для данных двух текстов, поскольку редко встречаются во всей коллекции, но по нескольку раз упоминаются в этих небольших документах.

Таким образом, TF-IDF демонстрирует наглядно, как делается вывод о схожести документов, т. е. данный алгоритм действительно отличается высокой степенью объяснимости.

**Пример 3.** Следующие два текста показали высокую степень схожести по алгоритму ARTM — 99 %.

- Американские исследователи синтезировали новые опиоидные анальгетики, которые лишены основных побочных эффектов этой группы препаратов. О результатах своей работы они сообщили в журнале Neuropharmacology. Сотрудники Университета Тулейн в Новом Орлеане модифицировали молекулы природных эндоморфинов (пептидных нейромедиаторов с высоким сродством к опиоидным мюрецепторам) и получили четыре новых пептида с продолжительным действием, которые способны проникать через гематоэнцефалический барьер. Эффекты полученных препаратов сравнили с морфином в серии экспериментов на крысах. Животным вводили дозы препаратов, эквивалентные по обезболивающему действию. Оказалось, что в отличие от морфина новые молекулы не вызывают существенного угнетения дыхания (основная причина смерти от передозировки опиоидов) и нарушения моторных функций. Кроме того, при длительном курсе лечения толерантность к ним развивается гораздо медленнее и в значительно меньшей степени. Стандартные тесты предпочтения места (животное находится дольше там, где получило наркотик) и самовведения (животное нажимает на рычаг, вводящий новую дозу препарата) показали, что новые препараты практически не вызывают привыкания и зависимости. По мнению исследователей, такая разница в побочных эффектах связана с действием препаратов на нейроглию (вспомогательные клетки нервной системы). Морфин активирует глиальные рецепторы p38, CGRP

и P2X7, что связывают с развитием толерантности к нему. У новых пептидов такого эффекта не обнаружено. Если подобная эффективность и безопасность полученных препаратов подтвердятся в клинических испытаниях, они станут новым "золотым стандартом" обезболивания, отмечают ученые. Приступить к исследованиям на людях планируется в ближайшие два года.

- Немецкие ученые обнаружили, что в бронхах человека присутствуют два типа обонятельных рецепторов, которые принимают участие в регуляции тонуса гладкой мускулатуры. Результаты работы опубликованы в журнале Frontiers in Physiology. Сотрудники Рурского университета в Бохуме с коллегами из других научных центров выделили гладкомышечные клетки из бронхиальной ткани, удаленной при оперативных вмешательствах. Анализ показал, что в этих клетках экспрессируются обонятельные рецепторы OR1D2 и OR2AG1, в виде как РНК, так и белка. После этого ученые выяснили, что при активации эти рецепторы повышают в клетках концентрацию кальция, регулируя их сократимость. Активация OR2AG1 амилбутиратом (эфиром с фруктовым запахом) предотвращала сокращение гладких мышц в ответ на стимуляцию гистамином — важным медиатором воспаления и аллергии. В свою очередь, активация OR1D2 бургеоналем (ароматическим альдегидом некоторых цветов) стимулировала мышечные сокращения (в живом организме это соответствует спазму бронхов). Кроме того, активация OR1D2-рецепторов вызывала выброс цитокинов: интерлейкина-8 и гранулоцитарно-макрофагального колониестимулирующего фактора. Оба эффекта этих рецепторов подавлялись специфическим антагонистом ундеканалем. Таким образом, в бронхах экспрессируются функциональные обонятельные рецепторы и принимают участие в регуляции физиологических и патологических процессов. Сократимость бронхиальной гладкой мускулатуры играет центральную роль в патогенезе астмы и других хронических воспалительных заболеваний дыхательной системы. Поэтому обнаруженные рецепторы могут стать перспективной мишенью для новых методов терапии этих болезней, пишут исследователи. В последнее десятилетие обонятельные рецепторы были обнаружены в различных тканях организма (до этого считалось, что они есть только у нейронов обонятельного анализатора). Так, исследования показали, что они регулируют восстановление кератиноцитов кожи, подвижность спермы, пролиферацию клеток рака простаты и гепатокарциномы.

Сходство этих двух текстов для человека менее очевидно, чем в предыдущем примере, поскольку нет такого же заметного количества совпадающих слов. И действительно, для данной пары алгоритм TF-IDF показывает весьма низкий результат: всего лишь 8 % схожести. И все-таки можно сразу сказать, что данные документы близки тематически: речь в обоих случаях идет о научных открытиях в области медицины.

Тем не менее сам алгоритм ARTM достаточно сложен для объяснения неподготовленному пользователю. Однако для обоснования сделанного вывода о схожести документов можно выделить наиболее выраженные в этих документах тематики, которые определяются наборами слов с некоторыми весами и вероятностями. Для данных двух текстов определяющей их сильную схожесть является тематика, связанная с медицинскими исследованиями и описанием физиологии организма, в том числе работы рецепторов. Стоит отметить, что значение степени близости (в данном случае 99 %) конкретной пары документов будет зависеть от общей тематической характеристики всего корпуса. В исследуемом корпусе "Тайга" представлены новостные заметки, которые освещают широкий спектр тем, поэтому новости об открытиях в сфере физиологии

оказываются тематически близкими на общем фоне. Однако если поместить те же два документа в корпус научно-медицинских статей, тематическая схожесть между ними уже будет меньше, поскольку статистика распределения медицинских понятий по документам корпуса изменится и алгоритм выделит другие, более узкие тематики.

Интересно, что, несмотря на совпадение большинства слов, согласно алгоритму ARTM первая пара документов схожа лишь на 80 %, и этот факт, с одной стороны, требует дополнительных усилий по объяснению, а с другой — иллюстрирует разницу между алгоритмами при измерении даже очевидной схожести текстов.

## Заключение

Машинное обучение помогает решать сложные задачи, связанные с выводами на основе больших массивов реальных данных. Для построения систем искусственного интеллекта такой подход наиболее уместен, когда знание неизвестно или скрыто. Однако для многих задач бизнеса и сценариев реальной жизни машинное обучение может применяться лишь при условии достаточно надежного обоснования результатов или объяснения путей их получения. Это особенно относится к тем случаям, когда решения могут иметь серьезные последствия.

Кроме того, такие области применения, как банковское дело, страхование и медицина, отличаются высокой степенью законодательного регулирования и требуют соблюдения всех установленных норм и правил. Однако не всегда этот массив знаний может быть своевременно изучен — в силу как своего большого объема, так и специфичности этой информации. Поэтому подобные прикладные знания должны быть своевременно, по мере необходимости, представлены в виде рекомендаций для пользователей ПИИС. Эта задача относится к области инженерии знаний. Инженерия знаний, с другой стороны, является целесообразной для представления экспертных знаний, о которых люди осведомлены и которые должны быть рассмотрены для соответствия требованиям или объяснениям.

Если информационные системы, которые делают знание явным, основаны на логике, их выводы легко могут быть объяснены. Такие системы обычно требуют бóльших первоначальных усилий в процессе разработки, чем те, что используют методы машинного обучения. Сократить усилия по разработке знаний позволяют такие перспективные подходы, как символическое и онтологическое обучение. Наконец, наблюдается растущий спрос на интеграцию инженерного знания и машинного обучения как взаимодополняющих компонентов интеллектуальной информационной системы. Недавние результаты показывают, что явно представленные прикладные знания могут помочь алгоритмам машинного обучения быстрее сходиться на разреженных данных и быть более устойчивыми к шуму. Однако инженерия данных требует использования методов с достаточной степенью объяснимости. Но при их выборе нужно также учитывать, какое смысловое сходство важнее для данной задачи: тематическая близость или совпадение слов.

Ранжирующие функции, выдающие понятные выводы о сходстве документов, не всегда подходят для решения бизнес-задач в рамках ПИИС, поскольку зачастую требуется обнаружить менее очевидные результаты, инсайты: найти схожие по смыслу документы, в которых нет явных совпадений.

Решающие эту задачу методы тематического моделирования трудно объяснимы, однако формальная метрика для оценки качества модели — когерентность тематик —

так же позволяет обеспечить достаточную степень очевидности тематического сходства между документами, выбранными алгоритмом. Таким образом, средняя когерентность тематик может служить косвенным критерием для оценки объяснимости метода: при определенном уровне когерентности полученных тематик решения системы о схожести между документами будут понятны пользователям.

В работе предложена формула для вычисления когерентности тематики, удобная с точки зрения объяснения. В ходе эксперимента установлено, что на выбранном корпусе когерентность достигает максимума при некотором количестве тематических компонентов, причем это оптимальное количество тематик различно для разных алгоритмов.

Эксперимент также продемонстрировал, что степень схожести не инвариантна относительно модели, т. е. некорректно сравнивать разные алгоритмы по одной и той же шкале, поскольку абсолютное значение схожести имеет неодинаковую важность для различных интеллектуальных алгоритмов. В эксперименте максимальная схожесть по TF-IDF составила 92 %, тогда как по LDA значение 93 % показали более 40 % документов из корпуса. Таким образом, в каждой прикладной задаче пороговое значение (например, для построения графа знаний) должно быть выбрано индивидуально с учетом особенностей конкретного алгоритма, выбранных параметров вычислений и характеристик исследуемого корпуса данных.

## Список литературы

[1] **Поспелов Д.А.** “Сознание”, “самосознание” и вычислительные машины. Системные исследования. Методологические проблемы. Ежегодник / Под ред. И.В. Блауберга, О.Я. Гельмана, В.П. Зинченко № 1, М.: Наука; 1969: 178–184.

[2] **Навроцкий М.А., Жукова Н.А., Муромцев Д.И., Мустафин Н.Г.** Методология проектирования, разработки и сопровождения доменных семантических порталов научно-технической информации. Науч.-техн. вестн. информ. технологий, механики и оптики. 2018; 18(2):286–298.

[3] **Голенков В.В., Гулякина Н.А., Давыденко И.Т., Шункевич Д.В., Еремеев А.П.** Онтологическое проектирование гибридных семантически совместимых интеллектуальных систем на основе смыслового представления знаний. Онтология проектирования. 2019; 9(1(31)):132–151.

[4] **Антонов В.В., Бармина Н.О., Никулина О.В.** Поддержка принятия решений при управлении программными проектами на основе нечеткой онтологии. Онтология проектирования. 2020; 10(1(35)):121–140.

[5] **Головко В.А., Голенков В.В., Ивашенко В.П., Таберко В.В., Иванюк Д.С., Крощенко А.А., Ковалев М.В.** Интеграция искусственных нейронных сетей с базами знаний. Онтология проектирования. 2018; 8(3(29)):366–386.

[6] **Minsky M., Kurzweil R., Mann S.** The society of intelligent veillance. IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life. 2013: 13–17.

[7] **Zanzotto F.M.** Viewpoint: Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research. 2019; (64): 243–252. DOI:10.1613/jair.1.11345.

[8] **Zhang S.** How to invest my time: Lessons from human-in-the-loop entity extraction. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019: 2305–2313.

[9] **Ng V., Rees E., Niu J., Zaghlool A., Ghiasbeglou H., Verster A.** Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. Canada Communicable Disease Report. 2020; 46(6):186–191.

[10] **Kluegl P., Toepfer M., Beck Ph.-D., Fette G., Puppe F.** UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering. 2016; 22(1): 1–40. DOI:10.1017/S1351324914000114.

[11] **Pauwels P., Zhang S., Lee Y.C.** Semantic web technologies in AEC industry: A literature overview. Automation in Construction. 2017; (73):145–165.

[12] **Azad H.K., Deepak A.** Query expansion techniques for information retrieval: A survey. Information Processing and Management. 2019; 56(5):1698–1735.

[13] **Bui Q.V., Sayadi K., Amor S.B., Bui M.** Combining Latent Dirichlet Allocation and K-means for documents clustering: effect of probabilistic based distance measures. Asian Conference on Intelligent Information and Database Systems. Springer, Cham; 2017: 248–257.

[14] **Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.** Neural architectures for named entity recognition. Proceedings of NAACL-HLT. 2016: 260–270. Available at: `https://www.aclweb.org/anthology/N16-1030.pdf`

[15] **McGovern A., Gagne D.J., Lagerquist R., Elmore K., Jergensen G.E.** Making the black box more transparent: Understanding the physical implications of machine learning. Bulletin of the American Meteorological Society. 2019; 100(11):2175–2199. DOI:10.1175/BAMS-D-18-0195.

[16] **Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L.** Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Turin, Italy: IEEE; 2018: 80–89. DOI:10.1109/DSAA.2018.00018.

[17] **Hagras H.** Toward human-understandable, explainable AI. Computer. 2018; 51(9):28–36.

[18] **Maaten L., Hinton G.** Visualizing data using t-SNE. Journal of Machine Learning Research. 2008; 9(11):2579–2605.

[19] **Curry H.B.** The method of steepest descent for non-linear minimization problems. Quarterly of Applied Mathematics. 1944; 2(3):258–261.

[20] **Sadeghi J., Sadeghi S., Niaki S.T.A.** Optimizing a hybrid vendor-managed inventory and transportation problem with fuzzy demand: An improved particle swarm optimization algorithm. Information Sciences. 2014; (272):126–144.

[21] **Kennedy J., Eberhart R.** Particle swarm optimization. Proceedings of ICNN'95 International Conference on Neural Networks. IEEE. 1995; (4):1942–1948.

[22] **Ribeiro M.T., Singh S., Guestrin C.** "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1135–1144. Available at: `https://www.aclweb.org/anthology/N16-3020.pdf`

[23] **Deerwester S., Dumais S., Furnas G.W., Landauer T.K., Harshman R.** Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990; 41(6):391–407.

[24] **Blei D.M., Ng A.Y., Jordan M.I.** Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; (3):993–1022.

[25] **Fevotte C., Idier J.** Algorithms for nonnegative matrix factorization with the $\beta$-divergence. Neural Computation. 2011; 23(9):2421–2456. Available at: `https://arxiv.org/pdf/1010.1763.pdf`

[26] **Vorontsov K., Potapenko A.** Additive regularization of topic models. Machine Learning. 2015; 101(1–3):303–323.

[27] **Huang G.** Supervised word mover's distance. Advances in Neural Information Processing Systems. 2016: 4862–4870.

[28] **Gomaa W.H., Fahmy A.A.** A survey of text similarity approaches. International Journal of Computer Applications. 2013; 68(13):13–18.

[29] **Levandowsky M., Winter D.** Distance between sets. Nature. 1971; 234(5323):34–35. DOI:10.1038/234034a0.

[30] **Salton G., Wu H.** A term weighting model based on utility theory. Proceedings of SIGIR. New York: ACM; 1980: 9—22.

[31] **Robertson S., Zaragoza H.** The probabilistic relevance framework: BM25 and beyond. Information Retrieval. 2009; 3(4):333–389. DOI:10.1561/1500000019.

[32] **Trotman A., Puurula A., Burgess B.** Improvements to BM25 and language models examined. Proceedings of the 2014 Australasian Document Computing Symposium. Melbourne, Australia; 2014: 58–65.

[33] **Shavrina T., Shapovalova O.** To the methodology of corpus construction for machine learning: "TAIGA" syntax tree corpus and parser. Trudy Mezhdunarodnoy Konferentsii "Korpusnaya Lingvistika — 2017". Saint-Peterburg: Izdatel'stvo SPbGU; 2017: 78–84.

[34] **Korobov M.** Morphological analyzer and generator for Russian and Ukrainian languages. International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham; 2015: 320–332.

[35] **Halko N., Martinsson P.G., Tropp J.A.** Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review. 2011; 53(2):217–288.

[36] **Hofmann T.** Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 50–57.

[37] **Newman D., Lau J.H., Grieser K., Baldwin T.** Automatic evaluation of topic coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California; 2010: 100–108.

[38] **Mimno D., Wallach H., Talley E., Leenders M., McCallum A.** Optimizing semantic coherence in topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK; 2011: 262–272.

Вычислительные технологии, 2020, том 25, № 5, с. 107–123. © ФИЦ ИВТ, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 5, pp. 107–123. © FRC ICT, 2020 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2020.25.5.009

## The explicability factor of the algorithm in the problems of searching for the similarity of text documents

Krasnov Fedor V.*, Smaznevich Irina S.

NAUMEN R&D, Ekaterinburg, Russia
*Corresponding author: Krasnov Fedor V., e-mail: fkrasnov@naumen.ru
*Received August 26, 2020, revised September 21, 2020, accepted September 25, 2020*

**Abstract**

The problem of providing a comprehensive explanation to any user why the applied intelligent information system suggests meaning similarity in certain texts imposes significant requirements on the intelligent algorithms. The article covers the entire set of technologies involved in the solution of the text clustering problem and several conclusions are stated thereof.

Matrix decomposition aimed at reducing the dimension of the vector representation of a corpus does not provide clear explanatiom of the algorithmic principles to a user. Ranking using the TF-IDF function and its modifications finds a few documents that are similar in meaning, however, this method is the easiest for users to comprehend, since algorithms of this type detect specific matching words in the compared texts. Topic modeling methods (LSI, LDA, ARTM) assign large similarity values to texts despite a few matching words, while a person can easily tell that the general subject of the texts is the same. Yet the explanation of how topic modeling works requires additional effort for interpretation of the detected ones. This interpretation gets easier as the model quality grows, while the quality can be optimized by its average coherence. The experiment demonstrated that the absolute value of documents similarity is not invariant for different intelligent algorithms, so the optimal threshold value of similarity must be set separately for each problem to be solved.

The results of the work can be further used to assess which of the various methods developed to detect meaning similarity in texts can be effectively implemented in applied information systems and to determine the optimal model parameters based on the solution explicability requirements.

*Keywords*: explainable artificial intelligence, XAI, ranking function, document similarity.

*Citation*: Krasnov F.V., Smaznevich I.S. The explicability factor of the algorithm in the problems of searching for the similarity of text documents. Computational Technologies. 2020; 25(5):107–123. DOI:10.25743/ICT.2020.25.5.009. (In Russ.)

**References**

1. **Pospelov D.A.** "Soznanie", "samosoznanie" i vychislitel'nye mashiny. Sistemnye issledovaniya. Metodologicheskie problemy. Ezhegodnik ["Consciousness", "self-awareness" and computers. System research. Methodological problems. Yearbook]. Pod red. I.V. Blauberga, O.Ya. Gel'mana, V.P. Zinchenko. Moscow: Nauka; 1969: 178–184. (In Russ.)
2. **Navrotskiy M.A., Zhukova N.A., Mouromtsev D.I., Mustafin N.G.** Design, development and maintenance methodology of domain semantic portals of scientific and technical information. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2018; 18(2):286—298.
3. **Golenkov V.V., Guliakina N.A., Davydenko I.T., Shunkevich D.V., Eremeev A.P.** Ontological design of hybrid semantically compatible intelligent systems based on sense representation of knowledge. Ontology of Designing. 2019; 9(1): 132–151.
4. **Antonov V.V., Barmina O.V., Nikulina N.O.** Decision-making support in software project management based on fuzzy ontology. Ontology of Designing. 2020; 10(1): 121–140.
5. **Golovko V.A., Golenkov V.V., Ivashenko V.P., Taberko V.V., Ivaniuk D.S, Kroshchanka A.A., Kovalev M.V.** Integration of artificial neural networks and knowledge bases. Ontology of Designing. 2018; 8(3): 366–386.
6. **Minsky M., Kurzweil R., Mann S.** The society of intelligent veillance. IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life. 2013: 13–17.
7. **Zanzotto F.M.** Viewpoint: Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research. 2019; (64): 243–252. DOI:10.1613/jair.1.11345.
8. **Zhang S.** How to invest my time: Lessons from human-in-the-loop entity extraction. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019: 2305–2313.
9. **Ng V., Rees E., Niu J., Zaghlool A., Ghiasbeglou H., Verster A.** Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. Canada Communicable Disease Report. 2020; 46(6):186–191.

10. **Kluegl P., Toepfer M., Beck Ph.-D., Fette G., Puppe F.** UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering. 2016; 22(1): 1–40. DOI:10.1017/S1351324914000114.
11. **Pauwels P., Zhang S., Lee Y.C.** Semantic web technologies in AEC industry: A literature overview. Automation in Construction. 2017; (73):145–165.
12. **Azad H.K., Deepak A.** Query expansion techniques for information retrieval: A survey. Information Processing and Management. 2019; 56(5):1698–1735.
13. **Bui Q.V., Sayadi K., Amor S.B., Bui M.** Combining Latent Dirichlet Allocation and K-means for documents clustering: effect of probabilistic based distance measures. Asian Conference on Intelligent Information and Database Systems. Springer, Cham; 2017: 248–257.
14. **Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.** Neural architectures for named entity recognition. Proceedings of NAACL-HLT. 2016: 260–270. Available at: `https://www.aclweb.org/anthology/N16-1030.pdf`
15. **McGovern A., Gagne D.J., Lagerquist R., Elmore K., Jergensen G.E.** Making the black box more transparent: Understanding the physical implications of machine learning. Bulletin of the American Meteorological Society. 2019; 100(11):2175–2199. DOI:10.1175/BAMS-D-18-0195.
16. **Gilpin L.H., Bau D., Yuan B.Z., Bajwa A., Specter M., Kagal L.** Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Turin, Italy: IEEE; 2018: 80–89. DOI:10.1109/DSAA.2018.00018.
17. **Hagras H.** Toward human-understandable, explainable AI. Computer. 2018; 51(9):28–36.
18. **Maaten L., Hinton G.** Visualizing data using t-SNE. Journal of Machine Learning Research. 2008; 9(11):2579–2605.
19. **Curry H.B.** The method of steepest descent for non-linear minimization problems. Quarterly of Applied Mathematics. 1944; 2(3):258–261.
20. **Sadeghi J., Sadeghi S., Niaki S.T.A.** Optimizing a hybrid vendor-managed inventory and transportation problem with fuzzy demand: An improved particle swarm optimization algorithm. Information Sciences. 2014; (272):126–144.
21. **Kennedy J., Eberhart R.** Particle swarm optimization. Proceedings of ICNN'95 International Conference on Neural Networks. IEEE. 1995; (4):1942–1948.
22. **Ribeiro M.T., Singh S., Guestrin C.** "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1135–1144. Available at: `https://www.aclweb.org/anthology/N16-3020.pdf`
23. **Deerwester S., Dumais S., Furnas G.W., Landauer T.K., Harshman R.** Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990; 41(6):391–407.
24. **Blei D.M., Ng A.Y., Jordan M.I.** Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; (3):993–1022.
25. **Fevotte C., Idier J.** Algorithms for nonnegative matrix factorization with the $\beta$-divergence. Neural Computation. 2011; 23(9):2421–2456. Available at: `https://arxiv.org/pdf/1010.1763.pdf`
26. **Vorontsov K., Potapenko A.** Additive regularization of topic models. Machine Learning. 2015; 101(1–3):303–323.
27. **Huang G.** Supervised word mover's distance. Advances in Neural Information Processing Systems. 2016: 4862–4870.
28. **Gomaa W.H., Fahmy A.A.** A survey of text similarity approaches. International Journal of Computer Applications. 2013; 68(13):13–18.
29. **Levandowsky M., Winter D.** Distance between sets. Nature. 1971; 234(5323):34–35. DOI:10.1038/234034a0.
30. **Salton G., Wu H.** A term weighting model based on utility theory. Proceedings of SIGIR. New York: ACM; 1980: 9—22.
31. **Robertson S., Zaragoza H.** The probabilistic relevance framework: BM25 and beyond. Information Retrieval. 2009; 3(4):333–389. DOI:10.1561/1500000019.
32. **Trotman A., Puurula A., Burgess B.** Improvements to BM25 and language models examined. Proceedings of the 2014 Australasian Document Computing Symposium. Melbourne, Australia; 2014: 58–65.

33. **Shavrina T., Shapovalova O.** To the methodology of corpus construction for machine learning: "TAIGA" syntax tree corpus and parser. Trudy Mezhdunarodnoy Konferentsii "Korpusnaya Lingvistika — 2017". Saint-Peterburg: Izdatel'stvo SPbGU; 2017: 78–84.
34. **Korobov M.** Morphological analyzer and generator for Russian and Ukrainian languages. International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham; 2015: 320–332.
35. **Halko N., Martinsson P.G., Tropp J.A.** Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review. 2011; 53(2):217–288.
36. **Hofmann T.** Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 50–57.
37. **Newman D., Lau J.H., Grieser K., Baldwin T.** Automatic evaluation of topic coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California; 2010: 100–108.
38. **Mimno D., Wallach H., Talley E., Leenders M., McCallum A.** Optimizing semantic coherence in topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK; 2011: 262–272.