ПРОГРАММНЫЙ МОДУЛЬ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

Павлюченко Кирилла Алексеевича
студент,
Белорусский государственный университет информатики и радиоэлектроники,
Республика Беларусь, г. Минск

Работа посвящена разработке метода для проверки авторства текста на русском языке. Предложенный метод включает анализ текста по метрике TF-IDF и по стилю письма. Описаны основные шаги алгоритма для обработки текстов, включая предварительную обработку, стемминг и метод анализа иерархий. Работа представляет интерес для людей, которые сталкиваются с плагиатом, в частности для преподавателей университетов.

Для учета важности критериев оценки авторства текста будет использован один из методов принятия решений, что позволит оценить принадлежность текста тому или иному автору.

Весь процесс начинается с того, что применяется операция стемминга, согласно которой выделяются основы слов. В результате получается текст, где все слова стеммированы и знаки препинания удалены [1]. Благодаря этой операции легче и удобнее работать с текстом.

Далее текст обрабатывается по метрике TF-IDF. TF-IDF – это мера важности слова в документе, которая учитывает как частоту встречаемости слова в документе, так и его частоту встречаемости во всех документах.

IDF вычисляется путем логарифмирования обратной частоты документа:

$$IDF(t) = \log_e \frac{a}{b} \quad (1)$$

где: a – общее количество документов в коллекции;

b – количество документов, в которых встречается терм t.

Если рассматривать IDF двух текстов, то это означает вычисление IDF для каждого слова, встречающегося в этих двух текстах. Как правило, в информационном поиске IDF используется для оценки важности слова в конкретном документе или запросе на поиск, поэтому для двух текстов IDF используется для сравнения важности различных слов в этих текстах. Если слово имеет более высокое значение IDF, то оно считается менее общим и более уникальным для данного набора документов, что может указывать на его большую важность в анализируемых текстах.

TF-IDF для каждого слова вычисляется путем перемножения TF на IDF [2]. Значение TF-IDF для каждого слова выражается в виде весового коэффициента, который показывает, насколько важно это слово в конкретном документе в контексте всей коллекции документов. Чем выше значение TF-IDF для конкретного слова в документе, тем более важным считается это слово в контексте документа.

К примеру, взяв два случайных текста на тему курения, можно составить следующую таблицу 1.

Таблица 1.

Расчет метрик

Words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
никотин	3/144=0,020	3/189=0,015	In (2/2)=0	0	0
провоцир	2/144=0,013	0	In (2/1)=0.693	0,009	0
котор	0,003	4/189=0,021	In (2/2)=0	0	0
рак	1/144=0,006	0	In (2/1)=0.693	0,004	0
причин	0	2/189=0,010	In (2/1)=0.693	0	0,006

Words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
привычк	0	1/189=0,005	In (2/1)=0.693	0	0,003
медикаментозн	0,006	0	In (2/1)=0.693	0,004	0
развит	0,013	0	In (2/1)=0.693	0,009	0
вред	0,02	0	In (2/1)=0.693	0,014	0
табачн	0	0,015	In (2/1)=0.693	0	0,011

Однако полученные значения довольно малы. Исходя их этого было принято решение, исправить полученные значения путем перемножения значений TF на 100 для увеличения веса данных показателей и повысить значения TF-IDF на 1 для того, чтобы нулевые значения имели вес.

Далее рассчитывается косинусная мера сходства между двумя документами на основе их TF-IDF векторов.

Косинусная метрика рассчитывается по формуле 2:

$$cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} \quad (2)$$

где: A и B – векторы TF-IDF для двух документов;

$||A||$ и $||B||$ – длины этих векторов.

Далее вычисляется средняя длина предложения в тексте. Для этого рассчитывается общее количество слов во всех предложениях. Затем общее количество слов делится на количество предложений. Эта длина нормализируется для дальнейшего использования.

Затем рассчитывается процент сложных предложений в тексте. Сначала текст разбивается на отдельные предложения. Затем каждое предложение проверяется на наличие "сложных союзов" (например, "что", "как", "где", "если" и так далее). Если предложение содержит хотя бы один из этих союзов, то оно считается сложным.

Количество сложных предложений и общее количество предложений подсчитывается, и затем вычисляется процент сложных предложений от общего количества предложений.

Далее с помощью метрики Евклида вычисляется схожесть двух векторов по средней длине предложения и проценту сложных предложений в тексте.

В контексте векторов, метрика Евклида используется для вычисления расстояния между двумя векторами в n-мерном пространстве. Расстояние между двумя векторами определено как длина линии, соединяющей две точки, которые представляют эти векторы в пространстве.

Чтобы вычислить расстояние между двумя векторами, используется формула:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3)$$

где: x_i, y_i – это расстояние между векторами X и Y.

Далее применяется метод Саати. Задаются функции полезности. Указываются критерии, по которым производится оценка состояний задачи:

– К1: метрика косинуса по TF-IDF;

– К2: метрика евклида по стилю текста.

Далее составляется матрица критериев (таблица 2).

Таблица 2.

Расчет метрик

	K1	K2
K1	1	1/3
K2	3	1

Затем находятся веса критериев: $\alpha_1 = 0{,}25, \alpha_2 = 0{,}75$.

Находим интегральную функцию полезности по формуле 4.

$$F = \sum_{i=1}^{n} \alpha_i \cdot \varphi(K_i), \quad (4)$$

где: φ – функция полезности по критерию.

По результатам вычисления Саати можно судить об авторе текста. Если результат вычисления меньше 0,5, то можно утверждать, что тексты написаны одним автором. Соответственно, если результат больше 0,5, то это означает, что тексты написаны разными авторами.

Список литературы:

1. Операция Стемминга [Электронный ресурс] – Электронные данные. – Режим доступа: https://coderlessons.com/tutorials/mashinnoe-obuchenie/uchebnik-nltk/5-stemming-i-lemmatizatsiia.
2. TF-IDF [Электронный ресурс] – Электронные данные. – Режим доступа: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/.