А. А. Алексеенко, А. В. Глазкова
Тюменский государственный университет, г. Тюмень
УДК 004.94

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ ПОДБОРА КЛЮЧЕВЫХ СЛОВ К НАУЧНЫМ ТЕКСТАМ

Аннотация. *В работе описывается принцип работы разработанного приложения для подбора ключевых слов к текстам научных работ на русском языке.*

Ключевые слова: *ключевые слова, Python, TF-IDF, TextRank, YAKE, корпус.*

Введение. В настоящее время алгоритмы и методы подбора ключевых слов уже протестированы для разных корпусов, однако ни одно исследование не проводилось на русскоязычных корпусах научных текстов. Для того, чтобы решить эту проблему, было решено сформировать русскоязычный корпус из научных работ, провести исследование эффективности алгоритмов и разработать приложение для подбора ключевых слов.

Методы. Для проведения исследования были взяты три алгоритма: TF-IDF, TextRank [1] и YAKE [2].

TF-IDF. Алгоритм TF-IDF основан на вычислении одноименной статистической метрики (TF-IDF), которая предназначена для обозначения важности отдельных слов для одного документа в коллекции документов и сортировки слов по важности для каждого документа.

Таким образом, можно выделить наиболее важные слова в тексте, которые могут быть кандидатами для включения в список ключевых слов.

TextRank. Это графовый алгоритм, который производит оценку важности слов в соответствии с их отношением в тексте. Ребра, построенные по алгоритму, являются ненаправленными взвешенными ребрами.

Основная формула алгоритма TextRank выглядит следующим образом:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \epsilon\, In(V_i)} \frac{\omega_{ji}}{\sum_{V_k \epsilon Out(V_j)} \omega_{jk}} WS(V_j), \quad (1)$$

где ω_{ji} используется для указания того, что пограничное соединение между двумя узлами имеет разные степени важности.

Конкретные шаги заключаются в следующем:

1. Разделить текст на предложения, $\mathrm{T} = [S_1, S_2, \dots, S_m]$.

2. Для каждого предложения $S_j \epsilon\, T$ удалить стоп-слова, оставить только существительные, глаголы и прилагательные.

3. Построить граф слов $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, где V — это набор узлов, состоящий из слов, сгенерированных на предыдущих этапах, и затем использующий отношение совместного появления слов для построения ребра между любыми двумя узлами: ребро существует между двумя узлами, только если их соответствующие слова совместно используются в n-грамме.

4. Согласно приведенной выше формуле, итеративно рассчитать вес каждого узла до сходимости.

5. Сортировать весов узлов в обратном порядке и получить из них t наиболее важных слов в качестве ключевых слов.

6. Для полученных t ключевых слов отметить их в исходном тексте, и, если между ними будут образованы смежные фразы, они будут извлечены как группы ключевых слов.

YAKE. Это графовый алгоритм, который производит отбор ключевых слов, основываясь на оценке их признаков. Основные шаги заключаются в следующем:

1. Проводится предобработка текста, которая заключается в разделении всего текста на отдельные слова.

2. Для каждого отдельного слова создается набор признаков, чтобы сохранить их особенные черты, такие как:

a. Регистр.

b. Позиция слова в тексте. Слова, которые встречаются в начале текста, имеют большую оценку, чем слова, которые находятся в середине текста.

c. Частота употребления слова. Чем чаще слово встречается в тексте, тем выше оценка по этому признаку.

d. Отношение слова к контексту в предложении. Чем больше разных слов встречается слева или справа от данного слова, тем меньше его значение для контекста.

e. Частота употребления слова в разных предложениях. Чем больше разных предложений содержит это слово, тем выше оценка по этому признаку.

3. Объединяем все оценки в одну меру, так что для каждого термина будет существовать оценка $S(w)$. Эта оценка послужит основой для процесса генерации ключевых слов, который должен быть выполнен на четвертом шаге.

На этом шаге мы рассматриваем скользящее окно из триграмм, тем самым создавая непрерывную последовательность из униграмм, биграмм и триграмм.

Каждой из них присваивается финальная оценка $S(kw)$, которая высчитывается по формуле:

$$S(kw) = \frac{\prod_{w \in kw} S(w)}{TF(kw) \times (1 + \sum_{w \in kw} S(w))}. \quad (2)$$

4. Очистка сгенерированных дубликатов.

5. Получаем список униграм, биграмм и триграмм с оценкой $S(kw)$, чем меньше оценка — тем больше важность слова.

Корпус. Для проведения испытаний был собран корпус из научных работ в области компьютерных технологий. В состав корпуса включены тексты научных работ, опубликованных в сборниках трудов конференций [3-7].

Корпус представляет собой JSON файл, в котором хранится множество словарей, соответствующих ключам «0», «1», «2» и т. д., каждый из которых содержит два ключа: «text» и «keywords», по которым можно получить соответственно: текст и оригинальные ключевые слова к этому тексту.

Характеристики собранного корпуса представлены в табл. 1.

Для сравнения качества извлечения ключевых слов с результатами, полученными для англоязычных корпусов, был взят корпус научных работ «Krapivin2009» [8], в котором также использованы научные работы из области компьютерных технологий.

Таблица 1

Характеристики собранного корпуса

Характеристика	*Значение*
Количество текстов	229
Среднее кол-во ключевых слов у текста	4,3
Максимальное кол-во ключевых слов у текста	11
Максимальная длина n-граммы	9
Максимальное количество символов в ключевом слове	78
Средняя длина n-граммы	1,9

Характеристики корпуса «Krapivin2009» представлены в табл. 2.

Таблица 2

Характеристики англоязычного корпуса

Характеристика	*Значение*
Количество текстов	229
Среднее кол-во ключевых слов у текста	5,34
Максимальное кол-во ключевых слов у текста	24
Максимальная длина n-граммы	8
Максимальное количество символов в ключевом слове	62
Средняя длина n-граммы	2,05

Результаты исследования. Для анализа результатов исследования использовались метрики ROUGE-1 и ROUGE-L [9], которые предназначены для сравнения автоматически сгенерированных ключевых слов с ключевыми словами, составленными человеком. Результаты исследования эффективности приведены в табл. 3.

Из результатов, приведенных в табл. 3, можно сделать вывод, что алгоритм YAKE при N = 2, показывает наилучший результат в 37,41%.

Затем были проведены аналогичные эксперименты с англоязычным корпусом, их результаты представлены в табл. 4.

Таблица 3

Результаты исследований алгоритмов на русскоязычном корпусе

Алгоритм	*ROUGE-l*	*ROUGE-L*
TF-IDF (N=1)*[1]	17,40%	15,35%
TextRank	23,81%	19,80%
YAKE (N=1)	35,45%	26,12%
YAKE (N=2)	37,41%	30,17%

Таблица 4

Результаты исследования алгоритмов на англоязычном корпусе

Алгоритм	*ROUGE-l*	*ROUGE-L*
TF-IDF (N=1)	18,29%	14,78%
TextRank	19,21%	15,70%
YAKE (N=1)	25,52%	20,37%
YAKE (N=2)	29,64%	24,63%

Из результатов, приведенных в таблице, можно сделать вывод, что алгоритм YAKE при N = 2, также стал наиболее эффективным для англоязычного корпуса.

На основании результатов исследования эффективности алгоритмов, для разработки приложения был выбран алгоритм YAKE.

Результаты разработки приложения. В результате работы было реализовано приложение для подбора ключевых слов на языке программирования Python, в основе которого используется алгоритм YAKE, а именно, его реализация на языке Python.

Приложение было реализовано в виде программы, которая позволяет пользователю ввести ссылку на файл формата PDF, и нажать кнопку «Подобрать ключевые слова», после чего в поле напротив надписи «Ключевые слова» появится список сгенерированных ключевых слов. Результат работы программы представлен на рис. 1.

[1] N — размер n-граммы.

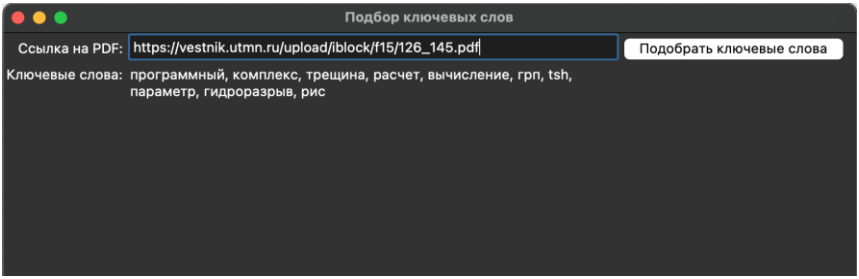

Рис. 1. Пример работы приложения

Заключение. В результате проведенной работы было проведено исследование эффективности трех алгоритмов подбора ключевых слов и разработано приложение для подбора ключевых слов на языке Python, в основу которого лег самый эффективный из всех трех алгоритмов, алгоритм YAKE.

Для взаимодействия с библиотекой, реализующей алгоритм YAKE, использовался Python.

Благодарности. Работа выполнена в рамках проекта № МК-3118.2022.4, реализованного за счет средств гранта Президента Российской Федерации для поддержки молодых ученых — кандидатов наук.

СПИСОК ЛИТЕРАТУРЫ

1. Mihalcea R. Textrank: Bringing order into text / R. Mihalcea, P. Tarau. — Text : direct // Proceedings of the 2004 conference on empirical methods in natural language processing, 25-26 July 2004. — Barcelona, 2004. — P. 404-411.
2. YAKE! Keyword extraction from single documents using multiple local features / R. Campos, V. Mangaravite, A. Pasquali [et al.]. — Text : direct // Information Sciences. — 2020. — Vol. 509. — P. 257-289.
3. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень: Издательство Тюменского государственного университета, 2020. — 735 с. — Текст : непосредственный.
4. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень: Издательство Тюменского государственного университета, 2019. — 376 с. — Текст : непосредственный.

5. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень : Издательство Тюменского государственного университета, 2018. — 496 с. — Текст : непосредственный.
6. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень : Издательство Тюменского государственного университета, 2017. — 534 с. — Текст : непосредственный.
7. Математика и междисциплинарные исследования : материалы Всероссийской научно-практической конференции молодых ученых с международным участием. — Пермь : Пермский государственный национальный исследовательский университет, 2020. — 343 с. — Текст : непосредственный.
8. Krapivin M. Large dataset for keyphrases extraction : Technical report / M. Krapivin, A. Autaeu, M. Marchese ; University of Trento. — Povo-Trento, 2008. — 6 p. — Text : direct.
9. Lin C. Y. Rouge: A package for automatic evaluation of summaries / C. Y. Lin. — Text : direct // Text summarization branches out. — Barcelona, 2004. — P. 74-81.

А. А. Кузьминых, А. А. Ступников
Тюменский государственный университет, г. Тюмень

УДК 004.912

РАЗРАБОТКА И АНАЛИЗ МОДЕЛЕЙ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ ДЛЯ ДАННЫХ ОБ УЧЕБНЫХ ДИСЦИПЛИНАХ ИМиКН

Аннотация. *В статье рассматривается подход к анализу соответствия содержания дисциплин, представленного в учебно-методических комплексах (УМК), определенному разделу компьютерных наук.*

Ключевые слова: *кластеризация, классификация, машинное обучение, анализ текстовых данных.*

Введение. Образовательные организации должны поддерживать высокий уровень качества профессиональной подготовки будущих квалифицированных специалистов. Одним из основных факторов образовательной деятельности является содержательное наполне-