УДК 004.738.5+331.101.1

ПРИЛОЖЕНИЕ ДЛЯ ГЕНЕРАЦИИ КЛЮЧЕВЫХ СЛОВ К ТЕКСТОВЫМ ДОКУМЕНТАМ

Марамыгина А.И.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Пилиневич Л. П. – д-р техн. наук, профессор, профессор кафедры ИПиЭ

Аннотация. Была разработана программа, осуществляющая генерацию ключевых слов для текстовых документов по их собственному содержанию. Для решения задачи были применены технологии машинного обучения и обработки естественных языков (*Natural Language Processing*). Обучение модели машинного обучения производилось на наборе текстовых данных, представляющем собой субтитры видеороликов с образовательного канала *TED*.

Ключевые слова: генерация ключевых слов, машинное обучение, обработка естественных языков

Введение. Скорость развития технологий в наш век рекордно большая для всей истории человечества. Мир движется к удобствам, экономии ресурсов, автоматизации и оптимизации всех процессов. В контексте информационных технологий во многом это удается благодаря разработке эффективных сервисов, которые направлены на облегчение работы людей, занимающихся рутинной и кропотливой деятельностью, которая, благодаря сервисам, теперь может выполняться практически без человеческого участия или контроля.

Программа генерации ключевых слов для текстовых документов может быть внедрена в самых разных областях, в которых производится работа с большими объёмами текстовых данных. Ключевые слова можно использовать в качестве инструмента для поиска информации в условиях обращения к большим объёмам текстовых данных, просмотреть и промаркировать которые вручную не представляется возможным или практичным. Кроме того, подобная программа может стать основой для последующей разработки продукта, позволяющего выделить в тексте главные мысли и резюмировать его содержание.

Автором статьи разработана программа, генерирующая ключевые слова с использованием алгоритмов машинного обучения (*ML*). Описаны технологии, примененные для подготовки и предварительной обработки данных.

Основная часть. Для создания программы генерации ключевых слов необходимо решить следующие задачи:

– определить наиболее подходящие технологии для разработки программы;
– получить качественные текстовые данные;
– очистить данные при необходимости;
– провести предобработку данных;
– осуществить векторизацию данных;
– обучить *ML*-модель, откалибровать её параметры.

Наиболее подходящим языком программирования для разработки программы генерации ключевых слов является язык *Python*, для которого существует большое число библиотек, предназначенных для работы с большими объемами данных.

Для получения качественных данных необходимо выбрать такой источник, который позволил бы обеспечить большие объёмы разнообразных по тематике и лексикону текстов. Анализ проводился среди англоязычных образовательных *YouTube*-каналов, в результате был выбран канал с записями конференций *TED*. Причиной выбора послужил объём загруженных видеороликов (более 3,500 на момент получения данных), наличие субтитров для практически всех видео, разнообразие тем выступлений и отсутствие рекламных интеграций, которые могли бы повлиять на качество работы алгоритмов.

После получения текстовых данных проводилась их очистка с помощью регулярных выражений. В качестве предобработки были удалены стоп-слова, регистр первых букв в предложениях изменён с верхнего на нижний. Была осуществлена лемматизация.

В качестве первой попытки сгенерировать ключевые слова была применена технология *TF-IDF* (*Term Frequency – Inverse Document Frequency*) – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса [1].

```
Out[13]: ([('supernova', 6.016755770621052),
           ('atom', 3.8321503245419897),
           ('star', 3.506738791982303),
           ('oxygen', 3.2485690110858716),
           ('explosion', 3.2465079603977616)],
          ['Astronomy',
           'Science',
           'Universe',
           'Human Origins',
           'Human Body',
           'Cosmos',
           'Humanity',
           'Visualizations',
           'Space',
           'Solar System'])
```

Рисунок 1 – Сравнение ключевых слов, предсказанных с помощью *TF-IDF*, с фактическими ключевыми словами

Для получения более качественных результатов была опробована технология *word2vec*. *Word2vec* – общее название для совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке [2, 3]. Так как все текстовые документы в наборе состоят из различного числа слов, а входные данные для обучения модели должны иметь постоянную размерность, необходимо было выбрать подход для обеспечения этого постоянства. Было решено выбрать самое значимое слово для каждого текста в соответствии с мерой *TF-IDF*, и его векторное представление в пространстве *word2vec* использовать в качестве набора входных признаков.

В качестве средства генерации выходных значений был обучен алгоритм *SVR* (*Support Vector Regression*).

Out[45]:

	y_predicted	y_true
41	NASA	Astronomy
42	science	Microbes
43	cancer	DNA
44	Physics	visible
45	biology	Science
46	poetry	culture
47	viruses	biology

Рисунок 2 – Сравнение ключевых слов, предсказанных с помощью *SVR* и *word2vec*, с фактическими ключевыми словами

Заключение. Была реализована программа генерации ключевых слов при помощи алгоритмов обработки естественных языков и машинного обучения, таких как лемматизация,

токенизация, *TF-IDF*, *SVR* и *word2vec*. Подход к выполнению задачи, описанный в статье, показал удовлетворительные результаты.

В качестве следующих шагов для улучшения работы программы следует улучшить обработку данных, а также провести генерацию ключевых слов, основываясь на более чем одном входном векторе. Возможно также применение рекуррентных нейронных сетей.

Список литературы

1. *Солтон Дж. Динамические библиотечно-поисковые системы. М.: — Мир, 1979. — 89 с.*
2. *Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. — 2013. — 193с.*
3. *Bengio Y., Ducharme R., Vincent P. A neural probabilistic language model // In Journal of Machine Learning Research. — 2003. 145 с.*

UDC 004.738.5+331.101.1

KEYWORD GENERATION APPLICATION FOR TEXT DOCUMENTS

Maramygina A.I.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Pilinevich L.P. – Dr. Tech. Sc., full professor, professor of the department of EPE

Annotation. A program was developed that generates keywords for text documents according to their content. Machine learning and Natural Language Processing technologies were used to solve the problem. The machine learning model was trained on a text dataset that consists of subtitles of videos from the TED educational channel.

Keywords: keywords generation, machine learning, natural language processing