УДК 004.047

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА УКРАИНОЯЗЫЧНЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

Голуб Татьяна

Запорожский национальный технический университет

Аннотация

В данной работе рассмотрены методы предварительной обработки текстовых документов. Предложен метод выделения украиноязычных текстов из совокупности текстов, представленных на английском, русском и украинском языках. А также представлен стеммер Портера для украиноязычных текстов.

Abstract

This paper discusses methods of text documents pre-treatment. It is offered the Ukrainian texts selection method of the submitted plurality texts in English, Russian and Ukrainian languages. Also represented Porter stemmer for Ukrainian-language texts.

Объем текстовой документации, которую специалист обрабатывает при поиске необходимой информации, существенен. Одним из способов упрощения задачи такой обработки является классификация этой документации. На данный момент в литературе предложено множество классификаторов текстовых документов, но, при своей универсальности, они не решают все проблемы. Так, наиболее распространенными анализируемыми документами в нашем регионе являются тексты, приведенные на английском и русском языках. Потому документы, представленные на украинском языке, чаще всего остаются без внимания. В данной работе рассматривается возможность предварительной подготовки текстовых документов источников, представленных на украинском языке параллельно с англоязычными и русскоязычными текстами.

Процесс классификации представляет собой несколько этапов: предварительная подготовка текста (создание вектора документа, обработка вектора с целью сокращения перечня слов и исключения неинформативных) и непосредственно классификация представленного в документе текста по заданным категориям [1].

Реализации приведенной последовательности действий обработки текста не являются универсальными вне зависимости от языка написания этого текста, так как разные языки имеют различные морфологические признаки. В связи с этим, первым этапом обработки текста становится определение использованного языка в том случае, если предполагается анализ многоязычных текстов. Предложенные в литературе методы определения языка текста предполагают статистический [2], семантический анализ текстов, либо дополнительное использование библиотек [3], что влечет за собой усложнение алгоритма принятия решения. Поскольку этап определения языка текста является промежуточным и он проводится на текстах русского, украинского либо английского языков, данные конструкции являются громоздкими и создают дополнительную нагрузку на вычислительную мощность оборудования, потому их нецелесообразно использовать для поставленной задачи. Следовательно, возникает потребность в оптимизации этого процесса.

Выделение из общего объема анализируемых документов текста, представленного на английском языке, автором предлагается выполнять путем проверки наличия букв латинского алфавита в доле, превышающей некое ранее заданное пороговое значение. Указанное пороговое значение компенсирует употребление букв латинского алфавита в текстах, написанных на русском и украинском языках. Задача дальнейшего выделения украинского языка вызывает проблемы в связи с его схожестью с русским языком. Альтернативным методом принятия решения предлагается проверка на наличие в словах

букв, свойственных конкретному языку. Так, для украинского языка такими буквами являются *і, ї, є, ґ* и апостроф, для русского – это *ё, ъ, ы э*. Но часто встречаются слова, не содержащие указанных признаков. В таком случае для идентификации рассматриваемых языков предлагается проверять также формы суффиксов и окончаний, которые являются специфичными в обоих случаях.

После определения языка текста следующим этапом является преобразование этого текста в векторную форму для последующего его сокращения. Сокращение вектора выполняется путем отсечения окончаний и суффиксов слов (стемминг), чтобы оставшаяся часть была одинакова для всех форм слова. На данный момент в литературе предложено множество стеммеров для разных языков. К таковым относятся стеммеры, основанные на алгоритмах с использованием подхода описаний правил усечения слов (стеммер Портера, Paice/Husk стиммер [4]) и основанные на лингвистических данных, использующих семантическую информацию слова (Y -стеммер [5], анализ К-грамм [6]).

Стеммер Портера, разработанный на основе описания правил отсечения окончаний и суффиксов, не требует какой-либо дополнительной базы для генерирования критериев усечения слов, что сокращает необходимый объем памяти для работы программного кода. Это позволяет создать автономную программу небольшого объема. Проблемы, возникающие при работе со стеммером данного вида (сильное, либо слабое усечение слов) не являются критичными при решении задачи классификации текста по тематикам в связи с обработкой форм встречающихся слов общим для всех текстов подходом. Не смотря на то, что украинский язык имеет общие признаки с русским, существующие отличия требуют самостоятельной разработки стеммера Портера для украинского языка. Данный стеммер, предложенный автором, позволяет обрабатывать тексты, представленные на украинском языке, и предоставлять их в виде, пригодном для дальнейшей предобработки в форме исключения неинформативных слов.

Таким образом, после выполнения указанных действий предварительной обработки текстовых документов для анализа украиноязычного текста, формируется список информативных частей слов, являющихся приемлемым для дальнейшей классификации документа с учетом использования конкретного языка. А выполнение данной предобработки предложенными автором методами позволит выделять и анализировать украиноязычный текст с использованием программного обеспечения, выдвигающего меньшие требования к вычислительным мощностям и памяти оборудования.

Список использованных источников:

1. Golub T., The Analysis of Text Documents Classifiers Constructing Methods / T. Golub // XIII International conference: Modern Problems Of Radio Engineering, Telecommunications, And Computer Science 2016, 23-26 February 2016. : тези доп., – Lviv-Slavsko, Ukraine, 2016. – P.742-745.

2. Chepovskiy Andrey, Language identification for texts written in transliteration / Andrey Chepovskiy, Sergey Gusev, Margarita Kurbatova – [Электронный ресурс], - Режим доступа: https://publications.hse.ru/chapters/72776618.

3. Vatanen Tommi,. Language identification of short text segments with n-gram models / Tommi Vatanen, Jaakko Vayrynen, Sami Virpioja // Proceedings of LREC, 2010, P. 3423-3430.

4. Moral Cristian, A survey of stemming algorithms in information retrieval / Cristian Moral, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez // Information research - vol. 19 no. 1, – 2014, p. 605-625.

5. Yatsko Viatcheslav, Y-stemmer / Yatsko Viatcheslav – Yatsko's Computational Linguistics Laboratory [Электронный ресурс], - Режим доступа: http://yatsko.zohosites.com/y-stemmer.html.

6. Lama Prabin, Clustering system based on text mining using the k-means algorithm / Prabin Lama – Bachelor's thesis (UAS) of Information Technology “Text Mining and Clustering 2013”, [Электронный ресурс], - Режим доступа: https://www.theseus.fi/bitstream/handle/10024/69505/Lama_Prabin.pdf?sequence=1.