МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра физики и аэрокосмических технологий

Аннотация к дипломной работе

АНАЛИЗ ОТКРЫТЫХ ДАННЫХ NASA МЕТОДАМИ NLP

Могучая Екатерина Игоревна

Научный руководитель — доцент Т.П. Янукович

Минск, 2022

Реферат

Дипломная работа состоит из введения, 3 разделов, заключения, списка используемой литературы, состоящего из 26 источников. Работа изложена на 57 листах печатного текста, содержит 29 рисунков и 2 таблиц.

Ключевые слова: машинное обучение, обработка естественного языка, анализ, данные, стемминг, лемматизация.

Целью данной работы является анализ данных NASA методами NLP с целью выявления наиболее часто встречающихся слов и выражений, а также проведения тематического моделирования.

Для выполнения данной цели были поставлены следующие задачи:

1. Изучение методов и алгоритмов обработки естественного языка.
2. Проведение анализа современных алгоритмов для решения проблемы компьютерного анализа естественных языков.
3. Проведение предобработки текста.
4. Проведение анализа данных NASA.

Актуальность проблематики данной дипломной работы обусловлена постоянным совершенствованием искусственного интеллекта и необходимостью совершенствовать понимание и обработку естественных языков. Важнейшим свойством естественных языков является их эволютивность, т.е. склонность к непрерывному развитию и трансформации. Это означает, что все, что связано с естественными языками является бесконечно-развивающимся процессом – прослеживается надобность модернизировать искусственные технологии на непрерывно изменяющийся процесс.

Также многозначность слов, значение которых зачастую зависит от контекста, усложняет задачу понимание естественного языка. В нейросетевых моделях смыслы слов и фраз представлены в форме векторов, что дает возможность производить сложные операции, которые и находятся автоматически в процессе обучения нейронных сетей. Прогресс в машинной обработке языка связан с постоянным расширением анализируемого машиной контекста.

Рэферат

Дыпломная праца складаецца з увядзення, 3 раздзелаў, заключэння, спісу выкарыстоўванай літаратуры, які складаецца з 26 крыніц. Праца выкладзена на 57 лістах друкаванага тэксту, змяшчае 29 малюнкаў і 2 табліц.

Ключавыя словы: машыннае навучанне, апрацоўка натуральнай мовы, аналіз, дадзеныя, стэмінг, лематызацыя.

Мэтай дадзенай працы з'яўляецца аналіз дадзеных NASA метадамі NLP з мэтай выяўлення найболей часта сустракаемых слоў і выразаў, а таксама правядзенні тэматычнага мадэлявання.

Для выканання гэтай мэты былі пастаўлены наступныя задачы:

1. Вывучэнне метадаў і алгарытмаў апрацоўкі натуральнай мовы.
2. Правядзенне аналізу сучасных алгарытмаў для вырашэння праблемы камп'ютарнага аналізу натуральных моў.
3. Правядзенне перадапрацоўкі тэксту.
4. Правядзенне аналізу дадзеных NASA.

Актуальнасць праблематыкі дадзенай дыпломнай працы абумоўлена пастаянным удасканаленнем штучнага інтэлекту і неабходнасцю ўдасканальваць разуменне і апрацоўку прыродазнаўчых моў. Найважнейшым уласцівасцю натуральных моў з'яўляецца іх эвалютыўнасць, г.зн. схільнасць да бесперапыннага развіцця і трансфармацыі. Гэта азначае, што ўсё, што звязана з натуральнымі мовамі з'яўляецца працэсам, які бясконца развіваецца – прасочваецца неабходнасць мадэрнізаваць штучныя тэхналогіі на працэс, які бесперапынна мяняецца.

Таксама шматзначнасць слоў, значэнне якіх часта залежыць ад кантэксту, ускладняе задачу разуменне натуральнай мовы. У нейросетевых мадэлях сэнсы слоў і фраз прадстаўлены ў форме вектараў, што дае магчымасць вырабляць складаныя аперацыі, якія і знаходзяцца аўтаматычна падчас навучанні нейронавых сетак. Прагрэс у машыннай апрацоўцы мовы злучаны са сталым пашырэннем аналізаванага машынай кантэксту.

Abstract

The thesis consists of an introduction, 3 chapters, a conclusion, a list of references, consisting of 26 sources. The work is presented on 57 sheets of printed text, contains 29 figures and 2 tables.

Keywords: machine learning, natural language processing, analysis, data, stemming, lemmatization.

The purpose of this work is to analyze NASA data using NLP methods in order to identify the most common words and expressions, as well as to conduct thematic modeling.

To achieve this goal, the following tasks were set:

1. The study of methods and algorithms for natural language processing.
2. Analysis of modern algorithms for solving the problem of computer analysis of natural languages.
3. Carrying out text preprocessing.
4. Analyzing NASA data.

The relevance of the problems of this thesis is due to the constant improvement of artificial intelligence and the need to improve the understanding and processing of natural languages. The most important property of natural languages is their volatility, i.e. propensity for continuous development and transformation. This means that everything related to natural languages is an endlessly evolving process - there is a need to upgrade artificial technologies to a continuously changing process.

Also, the ambiguity of words, the meaning of which often depends on the context, complicates the task of understanding natural language. In neural network models, the meanings of words and phrases are presented in the form of vectors, which makes it possible to perform complex operations, which are automatically in the process of training neural networks. Progress in machine language processing is associated with the constant expansion of the context parsed by the machine.