ҚАЗАҚСТАН РЕСПУБЛИКАСЫ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ

«Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ» КЕАҚ

Студенттер мен жас ғалымдардың
«ǴYLYM JÁNE BILIM - 2024»
XIX Халықаралық ғылыми конференциясының
БАЯНДАМАЛАР ЖИНАҒЫ

СБОРНИК МАТЕРИАЛОВ
XIX Международной научной конференции
студентов и молодых ученых
«ǴYLYM JÁNE BILIM - 2024»

PROCEEDINGS
of the XIX International Scientific Conference
for students and young scholars
«ǴYLYM JÁNE BILIM - 2024»

2024
Астана

УДК 001
ББК 72
G99

«ǴYLYM JÁNE BILIM – 2024» студенттер мен жас ғалымдардың XIX Халықаралық ғылыми конференциясы = XIX Международная научная конференция студентов и молодых ученых «ǴYLYM JÁNE BILIM – 2024» = The XIX International Scientific Conference for students and young scholars «ǴYLYM JÁNE BILIM – 2024». – Астана: – 7478 б. - қазақша, орысша, ағылшынша.

ISBN 978-601-7697-07-5

Жинаққа студенттердің, магистранттардың, докоранттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов имолодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 001
ББК 72
G99

ISBN 978-601-7697-07-5

The study found that using algorithms based on the TF-IDF metric, it is challenging to achieve high-quality clustering of textual information contained in short messages from the social network Twitter. It was concluded that the TF-IDF metric is not particularly suitable for clustering short text messages, or that a significant modification of this metric is necessary.

Algorithms based on «machine learning,» on the other hand, showed good results – six clusters of messages were identified: «study,» «emotions,» «photo sharing,» «urban environment,» «city news,» «politics.» This indicates a «rejuvenation" of the social network's audience. The data classification algorithm using the judging method is currently under development. In the course of further work, it is planned to compare the implemented text data classification algorithm and the LDA algorithm, as well as to optimize parallel clustering algorithms.

References

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

Vossen, G. (2014). Big data as the new enabler in business and other intelligence. Vietnam Journal of Computer Science, 1(1), 3-14.

Tamhane, D.S., & Sayyad, S.N. (2015). Big Data Analysis Using Hace Theorem. International Journal of Advanced Research (IJARCET), 4.

Tan, W., Blake, M. B., Saleh, I., & Dustdar, S. (2013). Social-network-sourced big data analytics. IEEE Internet Computing, (5), 62-69.

Vasilkov, A. (2015). How "big data" helps improve security. Retrieved September 24, 2015, from http://www.computerra.ru/108760/security-n-big-data/

Chubukova, I. Data Mining Tasks. Classification and clustering. -INTUIT.ru.

Blagov, A., Rytcarev, I., Strelkov, K., & Khotilin, M. (2015). Big Data Instruments for Social Media Analysis. Proceedings of the 5th International Workshop on Computer Science and Engineering, 179-184.

TF-IDF. (2015). Retrieved September 20, 2015, from https://ru.wikipedia.org/wiki/TF-IDF

Wang, H. (2014). Introduction to Word2vec and its application to find predominant word senses.

Yu, M., & Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In Association for Computational Linguistics (ACL), 545-550.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Blei, D.M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3, 993-1022.

УДК 004.048

АНАЛИЗАТОР ТЕКСТА В ВИДЕ ЧАТ-БОТА НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Абашев Арслан Азатабекович, Ажикенов Арман Русланович

arsik2005.2005@gmail.com

Студенты факультета информационных технологий, кафедры “Технологии Искусственного Интеллекта” ЕНУ им. Л. Н. Гумилева, Астана, Казахстан

Научный руководитель – Садвакасов Р.М.

Аннотация. Данная научная работа посвящена исследованию анализатора текста в форме чат-бота, основанного на технологиях искусственного интеллекта (ИИ). Работа охватывает введение в концепцию чат-ботов и их значимость в современном информационном обществе. Описывается технологическая основа анализатора текста, включая методы обработки естественного языка (Natural Language Processing, NLP), извлечения ключевых слов и анализа сентимента. Полученные результаты позволяют оценить эффективность и потенциал использования анализатора текста в виде чат-бота для обработки и анализа текстовой информации в реальном времени.

Ключевые слова. Анализатор текста, NLP, чат-бот, тональность.

Введение. В современном мире активно расширяются сферы обмена информацией через Интернет. Рост количества социальных сетей, блогов, форумов и веб-ресурсов ставит перед нами задачу как анализ текстов, оставленных пользователями на различные темы: от их реакции на события до отзывов о товарах и услугах, а также мнений о высказываниях и оценок других.

Также с огромной скоростью развиваются и чат-боты, основанные на искусственном интеллекте. Они представляют собой программные агенты, способные анализировать и обрабатывать текстовую и любую другу информацию. В данной работе рассматривается реализация анализатора текста в форме чат-бота, который использует передовые методы искусственного интеллекта для взаимодействия с пользователями и предоставления ответов на их запросы [1].

Преимущества анализатора текста:

- Эффективность анализа текста: определение эмоциональной окраски (сентимента), извлечение ключевых слов и подсчет слов и предложений.
- Использование передовых технологий: Использование библиотеки Natural Language Toolkit (NLTK);
- Автоматизация процесса анализа: позволяет значительно ускорить и упростить этапы обработки текстовых данных.
- Учет контекста и семантики: повышает качество и надежность результатов анализа, что важно для принятия информированных решений на основе текстовых данных.

Методы проведенных исследований:

Введение в анализатор текста и его технологическую основу. Данная глава направлена на рассмотрение анализатора текста и его ключевых технологических аспектов. В современном информационном обществе анализ текста играет важную роль в обработке и понимании огромного объема информации, создаваемой и потребляемой пользователями Интернета. Анализатор текста позволяет автоматически обрабатывать текстовые данные, извлекать из них полезную информацию и применять ее в различных сферах, таких как маркетинг, социология, медицина, финансы и другие.

Одной из ключевых концепций, на которых основан анализ текста, является обработка естественного языка (Natural Language Processing, NLP). NLP позволяет компьютерам понимать, интерпретировать и генерировать человеческий язык, что позволяет им анализировать текстовую информацию на уровне, понятном человеку. Это включает в себя задачи такие как определение частей речи, выявление синтаксических структур, анализ семантики и контекста текста.

Другой важной технологией, используемой в анализе текста, являются методы извлечения ключевых слов. Эти методы позволяют автоматически определять наиболее важные и релевантные слова или фразы в тексте, что помогает сжимать информацию и выделять наиболее важные аспекты текста для последующего анализа [2].

Кроме того, анализ сентимента является еще одним важным аспектом анализа текста, который позволяет определять тональность высказываний. Это может быть полезно для оценки отзывов о продуктах или услугах, мониторинга общественного мнения или анализа тональности текстов в социальных сетях [3].

Разработка и реализация анализатора текста. После формулирования требований начинается выбор подходящих технологий и инструментов для реализации анализатора текста. Решение обычно зависит от требований проекта. При выборе технологий учитывается их способность обеспечивать высокую производительность, надежность и масштабируемость системы (рисунок - 1)

```
import requests
from tkinter import simpledialog, messagebox, Text, Scrollbar
import asyncio
import tkinter as tk
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.sentiment import SentimentIntensityAnalyzer
from collections import Counter
from nltk.tokenize import RegexpTokenizer
from nltk.stem import PorterStemmer
```

Рисунок 1-Список используемых библиотек

После выбора технологий разрабатывается архитектура анализатора текста. Это включает в себя определение модулей системы, их взаимосвязей и функциональных возможностей. Важно обеспечить гибкость и расширяемость архитектуры, чтобы в будущем можно было легко добавлять новые функции и улучшать систему. Затем происходит непосредственная реализация анализатора текста, включая написание кода, разработку пользовательского (рисунок - 2) [4].

```
def analyze_sentiment(self, text):
    sia = SentimentIntensityAnalyzer()
    scores = sia.polarity_scores(text)
    return scores

1 usage
def extract_keywords(self, text):
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(text)

    tokens = self.setting(tokens)
    filtered_tokens = tokens

    porter = PorterStemmer()
    stemmed_tokens = [porter.stem(word) for word in filtered_tokens]

    word_freq = Counter(stemmed_tokens)

    num_keywords = min(5, len(word_freq))
    keywords = word_freq.most_common(num_keywords)

    return [word for word, _ in keywords]

1 usage
def setting(self, tokens):
    stop_words = set(stopwords.words('english'))
    stop_words.update(set(stopwords.words('russian')))
    stop_words.update(set(stopwords.words('kazakh')))
    return [word.lower() for word in tokens if word.lower() not in stop_words]


1 usage
def count_words_and_sentences(self, text):
    words = word_tokenize(text)
    words_without_punctuation = [word for word in words if word.isalpha()]
    sentences = nltk.sent_tokenize(text)
    return len(words_without_punctuation), len(sentences)
```

Рисунок 2 -Написание алгоритма анализа

После завершения реализации проводится тестирование анализатора текста. Наконец, после успешного прохождения всех тестов система готова к оптимизации и внедрению. Оптимизация может включать в себя улучшение алгоритмов, оптимизацию базы данных и настройку конфигурации серверов.

Список использованных источников

1. Половнева М.В. Анализ развития и применения чат-бота.

[Электронный ресурс]: - Режим доступа: https://cyberleninka.ru/article/n/analiz-razvitiya-i-primeneniya-tehnologii-chat-bot/viewer. 2022. УДК 004.896.

2. Столбун Е. А., Полоско Е. И., Искусственный интеллект и обработка естественного языка. Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь.

[Электронный ресурс]: - Режим доступа: https://libeldoc.bsuir.by/bitstream/123456789/52258/1/Stolbun_Iskusstvennii.pdf. 2023. Стр. 1 - 3

3. Семина Т.А. Анализ тональности текста: современные подходы и существующие проблемы.

[Электронный ресурс]: - Режим доступа: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-sovremennye-podhody-i-suschestvuyuschie-problemy/viewer. 2020. УДК 04.008

4. Митина О.В., Евдокименко А.С., Методы анализа текста: методологические основания и программная реализация.

[Электронный ресурс]: - Режим доступа: https://cyberleninka.ru/article/n/metody-analiza-teksta-metodologicheskie-osnovaniya-i-programmnaya-realizatsiya/viewer. 2010. УДК 159.98

УДК 159.98

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МЕТОДОВ СОЗДАНИЯ ТЕМАТИЧЕСКИХ СЛОВАРЕЙ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Аманкелдин Акжол Медетулы
Akzhol.amankeldin@inbox.ru
Магистрант 2 курса Казахского национального университета имени Аль-Фараби,
Факультет информационных технологий, компьютерная лингвистика, Алматы, Казахстан
Научный руководитель - Рахимова Д.Р

Аннотация. В данной статье исследуется процесс создания тематических словарей на основе методов машинного обучения. Тематические словари представляют собой важный инструмент для анализа и классификации текстовой информации по различным тематикам. С развитием технологий и увеличением объемов данных становится необходимым использование эффективных методов обработки и анализа текста. Машинное обучение предлагает мощные инструменты для автоматизации процесса создания таких словарей, позволяя выявлять ключевые слова и понятия, характеризующие конкретные темы.

Введение

В статье рассматриваются основные этапы создания тематических словарей, включая сбор и предварительную обработку данных, выбор методов машинного обучения, а также оценку результатов. Приводятся примеры распространенных методов машинного обучения, применяемых для создания тематических словарей, таких как кластеризация, тематическое моделирование и использование нейронных сетей.

Также обсуждаются преимущества использования методов машинного обучения для создания тематических словарей, такие как повышение точности и автоматизация процесса. Приводятся примеры исследований, демонстрирующих успешное применение машинного обучения в данной области.

Наконец, подчеркивается важность дальнейших исследований для совершенствования методов создания тематических словарей на основе машинного обучения, особенно в контексте постоянно растущего объема и разнообразия текстовой информации.

Создание Тематических Словарей на Основе Машинного Обучения

С появлением больших объемов данных и расширением области применения машинного обучения возникла потребность в эффективных методах обработки и анализа текстов. Одной из ключевых задач в этой области является создание тематических словарей,