ИСПОЛЬЗОВАНИЕ БИБЛИОТЕКИ «NLTK» ПРИ РАБОТЕ С ТЕКСТАМИ НА РУССКОМ ЯЗЫКЕ

Яхья-заде А.С.

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

Петрова Н.Е. – к.филол.н., доцент

Данная статья рассматривает использование библиотеки Natural Language Toolkit при анализе текстов на русском языке. В работе обсуждаются основные функциональные возможности библиотеки, включая токенизацию, лемматизацию, стемминг, анализ синтаксиса и частей речи, а также применение методов машинного обучения для классификации текстов.

Одним из сложнейших аспектов искусственного интеллекта (ИИ), несомненно, является естественный язык. Долгое время его анализ и обработка оставались предметом исследований, требующих глубокого понимания языка, контекста и культурных нюансов. В связи с этим, появление NLTK (Natural Language Toolkit) – библиотеки для работы с естественным языком в Python – было значимым событием в области компьютерной лингвистики и искусственного интеллекта.

Библиотека NLTK была создана в начале 2000-х годов как проект по обработке естественного языка в рамках программы для обучения и исследований в Университете Пенсильвании [1]. Стремление к созданию структурированного и удобного инструмента для анализа текстовых данных вдохновило разработчиков на создание этой библиотеки. Они стремились не только сделать её эффективной, но и обеспечить доступность для широкого круга пользователей.

В настоящее время информация является одним из основных ресурсов, с которым мы сталкиваемся ежедневно. Обработка текстов становится все более важной задачей в различных областях, таких как анализ данных, машинное обучение, компьютерная лингвистика и другие. На сегодняшний день библиотека Natural

Language Toolkit (NLTK) является одним из наиболее популярных инструментов для обработки текстовых данных на различных языках, включая русский [2].

NLTK предоставляет широкий спектр функций для работы с текстами, включая токенизацию, лемматизацию, стемминг, анализ синтаксиса, работу с частями речи, а также возможности для обучения и применения моделей машинного обучения [3]. Данная библиотека позволяет проводить анализ текстов на этом языке с высокой точностью и эффективностью. Одним из основных преимуществ библиотеки NLTK является её гибкость и наличие большого количества инструментов, которые могут быть легко интегрированы в проекты обработки текста на русском языке. Остановимся подробнее на некоторых возможностях этого продукта.

Одной из ключевых функций библиотеки NLTK является токенизация (англ. *tokenization*). Библиотека предоставляет инструменты для разделения текста на отдельные слова или фразы, что является первым шагом в анализе текста. Мы считаем, что для русского языка это очень важно из-за его специфических особенностей, таких как использование сложных словоформ и пунктуации.

Также важным этапом при анализе текста на русском языке является процесс приведения слов к своей базовой форме, что обусловлено богатой морфологией. Для этого NLTK предоставляет возможность лемматизации (англ. *lemmatization*). NLTK в особенности предоставляет возможность лемматизации русских слов, что значительно упрощает анализ текстов.

NLTK также поддерживает стемминг (англ. *stemming*), который аналогичен лемматизации, но более грубый, приводящий слова к их основе, так называемому «стему». Хотя стемминг не так точен, как лемматизация, он может быть полезен в некоторых задачах обработки текста.

Кроме того, библиотека NLTK предоставляет инструменты для анализа синтаксиса предложений на русском языке, а также определения частей речи слов (англ. *Syntax Parsing and Part-of-Speech Tagging*). Это полезно для понимания структуры предложений и выделения ключевой информации из текста. NLTK также предоставляет возможности для обучения моделей машинного обучения на текстовых данных, что позволяет автоматизировать процессы анализа и классификации текстов на русском языке.

По нашему мнению, использование NLTK для анализа русскоязычных текстов позволяет исследователям и разработчикам эффективно извлекать информацию из больших объемов данных. Помимо основных функций, NLTK также предоставляет возможность работы с корпусами текстов [4], что облегчает обучение и тестирование алгоритмов на реальных данных. Одним из ключевых преимуществ NLTK является его открытый исходный код [5], что позволяет разработчикам модифицировать и дорабатывать функционал библиотеки под свои потребности.

При решении задач обработки естественного языка на русском для достижения более высокой точности и производительности NLTK может использоваться в сочетании с другими инструментами и библиотеками, такими как *spaCy* или *TensorFlow* [6].

Библиотека spaCy явялется инструментом для обработки естественного языка (англ. *Natural Language Processing*, *NLP*), который предоставляет простой и эффективный интерфейс для выполнения различных задач обработки текста на разных языках, включая русский. Среди функциональности spaCy можно выделить токенизацию, лемматизацию, выделение именованных сущностей, синтаксический анализ и др. TensorFlow – это открытая платформа для машинного обучения, разработанная командой *Google Brain*. Она предоставляет обширный инструментарий для создания и обучения различных моделей искусственных нейронных сетей.

Помимо этого, библиотека NLTK также предлагает предобученные модели для различных языков, включая русский. Библиотека обеспечивает доступ к различным методам векторизации текста, включая *TF-IDF* и *Word2Vec*, что позволяет представить текстовые данные в виде числовых векторов для последующего анализа и обработки. Важным аспектом использования NLTK является его активное сообщество пользователей и разработчиков, что способствует обмену знаний и опытом в области обработки текстов на русском языке.

Библиотека NLTK также предоставляет функционал для работы с различными форматами текстовых данных, включая файлы в форматах *TXT*, *CSV*, *XML* и другие, что делает его удобным инструментом для работы с разнообразными источниками информации [7]. Важно отметить, что NLTK имеет документацию и обучающие материалы на русском языке, что делает его доступным для широкого круга пользователей, включая русскоязычных исследователей и разработчиков. Необходимо также учитывать, что, хотя NLTK является мощным инструментом для анализа текстов на русском языке, его использование требует некоторой экспертизы в области обработки естественного языка и программирования.

Таким образом, использование NLTK при работе с текстами на русском языке упрощает и автоматизирует множество задач обработки текста, делая процесс анализа более эффективным и точным. Благодаря своей гибкости и мощным функциональным возможностям, NLTK остается одним из наиболее популярных инструментов в области обработки текстовых данных на русском языке.

Список использованных источников:

1. Natural Language Toolkit [Электронный ресурс]. – Режим доступа: https://deepgram.com/ai-glossary/natural-language-toolkit-nltk. – Дата доступа: 03.04.2024.

2. Руководство по NLTK с использованием Python [Электронный ресурс]. – Режим доступа: https://datafinder.ru/products/rukovodstvo-po-nltk-s-ispolzovaniem-python. – Дата доступа: 30.03.2024.

3. Обработка естественного языка с использованием Python [Электронный ресурс]. – Режим доступа: https://www.nltk.org/book/ch01.html. – Дата доступа: 30.03.2024.

4. Обработка естественного языка [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wlkl/index.php. – Дата доступа: 30.03.2024.

5. Исходный код библиотеки [Электронный ресурс]. – Режим доступа: https://github.com/nltk/nltk. – Дата доступа: 31.03.2024.

6. SpaCy против TensorFlow [Электронный ресурс]. – Режим доступа: https://stackshare-io.translate.goog/stackups/spacy-vs-tensorflow?_x_tr_sl=en&_x_tr_tl=ru&_x_tr_hl=ru&_x_tr_pto=sc. – Дата доступа: 31.03.2024.

7. Пример использования данных [Электронный ресурс]. – Режим доступа: https://www.nltk.org/howto/data.html. – Дата доступа: 31.03.2024.