УДК: 004.67

СИСТЕМЫ ТЕКСТОВОГО ПОИСКА ОБРАБОТКИ И АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА

♦♦♦♦♦

TEXT SEARCH PROCESSING AND NATURAL LANGUAGE ANALYSIS

Янаева М.В.
кандидат технических наук, доцент
кафедры информационные системы и программирование,
Кубанский государственный технологический институт

Ковалев Н.С.
студент 4 курса,
факультет «ИКСиИБ»
Кубанский государственный технологический институт
kovalov_1998@mail.ru

Боярко А.Э.
студент 4 курса,
факультет «Прикладная информатика» института
компьютерных систем и информационной безопасности
merts1411@mail.ru

Аннотация. В настоящее время жизнь человека окружена всевозможными сервисами, они просто неотъемлемая часть повседневной жизни. Сложно не заметить колоссальный темп роста систем поиска и обработки естественного языка. Все эти системы и виртуальные помощники, проникающие в наши дома и сердца, имеют в своей основе поисковую технологию. Каждый раз, когда вы запрашиваете систему, базу данных или приложение, эта система должна решить, какие результаты отображать или сказать – это поисковое приложение. OpenTable, Tinder, Siri, Алиса – все это поисковые приложения и виртуальные помощники. Технология поиска лежит в основе почти каждого популярного программного приложения, которое вы используете сегодня на работе, дома, в играх, за рабочим столом или на смартфоне.

Ключевые слова: обработка естественного языка, NLP, анализ NLP, библиотеки Python, Java.

Yanaeva M.V.
Candidate of Technical Sciences,
Associate Professor
of the Department of Information
Systems and Programming,
Kuban State Technological Institute

Kovalev N.S.
4th year Student,
Faculty of Applied Informatics
Institute of Computer Systems
and Information Security,
Kuban State Technological Institute
kovalov_1998@mail.ru

Boyarko A.E.
4th year Student,
Faculty of Applied Informatics
Institute of Computer Systems
and Information Security,
Kuban State Technological Institute
merts1411@mail.ru

Annotation. Currently, human life is surrounded by all sorts of services, they are just an integral part of everyday life. It's hard not to notice the tremendous growth of natural language search and processing systems. All these systems and virtual assistants that penetrate our homes and hearts are based on search technology. Every time you query a system, database, or application, that system has to decide what results to display or say – it's a search application. OpenTable, Tinder, Siri, Alice are all search apps and virtual assistants.Search technology is at the heart of almost every popular software application you use today at work, at home, in games, at your desktop or on your smartphone.

Keywords: natural language processing, NLP, NLP analysis, Python libraries, Java.

Обработка естественного языка (NLP), технология, которая обеспечивает работу всех чат-ботов, голосовых помощников, интеллектуального текста и других речевых / текстовых приложений, которые пронизывают нашу жизнь, значительно изменилась за последние несколько лет [1]. Существует большое разнообразие инструментов NLP.

Natural Language Toolkit (NLTK)

Было бы легко утверждать, что Natural Language Toolkit (NLTK) является наиболее полнофункциональным инструментом [2]. Он реализует практически любой компонент NLP, который вам понадобится, например, классификацию, токенизацию, стемминг, маркировку, синтаксический анализ и семантические рассуждения. И часто существует более одной реализации для каждой, так что вы можете выбрать точный алгоритм или методологию, которые вы хотели бы использовать. Он также поддерживает

множество языков. Однако он представляет все данные в виде строк, что хорошо для простых конструкций, но затрудняет использование некоторых дополнительных функций. Документация также довольно плотная, но ее много, как и отличной книги. Библиотека также немного медленная по сравнению с другими инструментами. В целом, это отличный инструментарий для экспериментов, исследований и приложений, которым требуется определенная комбинация алгоритмов.

SpaCy

SpaCy, вероятно, является основным конкурентом NLTK. В большинстве случаев он быстрее, но имеет только одну реализацию для каждого компонента NLP. Кроме того, он представляет все в виде объекта, а не строки, что упрощает интерфейс для создания приложений. Это также помогает ит интегрироваться со многими другими фреймворками и инструментами обработки данных, так что вы сможете делать больше, когда лучше разберетесь в своих текстовых данных. Однако SpaCy не поддерживает столько языков, сколько NLTK. У него простой интерфейс с упрощенным набором опций и отличной документацией, а также множество нейронных моделей для различных компонентов обработки и анализа языка. В целом, это отличный инструмент для новых приложений, которые должны быть производительными в рабочей среде и не требуют специального алгоритма.

Инструменты Java [3]

OpenNLP

OpenNLP размещается ApacheFoundation, поэтому его легко интегрировать в другие проекты Apache, такие как ApacheFlink, ApacheNiFi и ApacheSpark. Это общий инструмент NLP, который охватывает все распространенные компоненты обработки NLP, и его можно использовать из командной строки или в приложении в качестве библиотеки. Он также имеет широкую поддержку нескольких языков. В целом, OpenNLP– это мощный инструмент с множеством функций, готовый к рабочей нагрузке, если вы используете Java.

StanfordNLP

StanfordCoreNLP – это набор инструментов, который обеспечивает статистическое НЛП, НЛП глубокого обучения и функциональность НЛП на основе правил. Было создано множество других привязок к языкам программирования, чтобы этот инструмент можно было использовать за пределами Java. Это очень мощный инструмент, созданный элитным исследовательским учреждением, но, возможно, он не самый лучший для производственных нагрузок. Этот инструмент имеет двойную лицензию со специальной лицензией для коммерческих целей. В целом, это отличный инструмент для исследований и экспериментов, но это может повлечь за собой дополнительные затраты в производственной системе. Реализация Python также может заинтересовать многих читателей больше, чем версия Java. Кроме того, один из лучших курсов машинного обучения преподается профессором Стэнфорда на Coursera. Проверьте это вместе с другими замечательными ресурсами.

AmazonComprehend – это сервис NLP, интегрированный с инфраструктурой AmazonWebServices. Вы можете использовать этот API для задач NLP, таких как анализ настроений, тематическое моделирование, распознавание сущностей и многое другое.

Для тех, кто работает в сфере здравоохранения, есть специализированный вариант: AmazonComprehendMedical, который позволяет выполнять расширенный анализ медицинских данных с использованием машинного обучения.

Gensim – это узкоспециализированная библиотека Python, которая в основном занимается задачами тематического моделирования с использованием таких алгоритмов, как скрытое распределение Дирихле (LDA). Он также отлично подходит для распознавания сходства текстов, индексирования текстов и навигации по различным документам.

Инструменты обработки естественного языка помогают компаниям получать информацию из неструктурированных текстовых данных, таких как электронные письма, онлайн-обзоры, сообщения в социальных сетях и многое другое.

Существует множество онлайн-инструментов, которые делают NLP доступным для вашего бизнеса, например, с открытым исходным кодом и SaaS. Библиотеки с открытым исходным кодом бесплатны, гибки и позволяют разработчикам полностью настраивать их. Однако они не являются экономически эффективными, и вам нужно будет потратить время на создание и обучение инструментов с открытым исходным кодом, прежде чем вы сможете воспользоваться преимуществами.

Литература

1. Нейролингвистическое программирование [Электронный ресурс]. – URL : https://ru.wikipedia.org/wiki/(дата обращения: 22.08.2022)
2. Witkowski, Tomasz. Thirty-Five Years of Research on Neuro-Linguistic Programming. NLP Research Data Base.State of the Art or Pseudoscientific Decoration?(англ.) // PolishPsychological-Bulletin: journal. – 2010. – doi : 10.2478/v10059-010-0008-0. (дата обращения: 22.08.2022.
3. Обработка естественного языка на Java / пер. сангл. А.В. Снастина. – М. : ДМК Пресс, 2016. (дата обращения: 22.08.2022)

References

1. Neuro-Linguistic Programming [Electronic resource]. – URL : https://ru.wikipedia.org/wiki/ (date of access: 22.08.2022)
2. Witkowski, Tomasz. Thirty-Five Years of Research on Neuro-Linguistic Programming. NLP Research Data Base.State of the Art or Pseudoscientific Decoration?(English) // PolishPsychologicalBulletin: journal. – 2010. – doi : 10.2478/v10059-010-0008-0. (Date of access: 22.08.2022.
3. Natural language processing in Java / per. English A.V. Snap. – M. : DMK Press, 2016. (Date of access: 22.08.2022)