Петров Александр Васильевич,
студент,
«ЧГУ им. И.Н. Ульянова»,
г. Чебоксары, Чувашская республика

РАЗРАБОТКА АРХИТЕКТУРЫ ЛЕКСИЧЕСКОГО ПОИСКОВИКА ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА НА ЯЗЫКЕ JAVA

Аннотация. В статье детально описывается архитектура и функционал поисковика для национального корпуса чувашского языка, с учетом всех требований и особенностей. Поисковик представляет собой специализированную программу для анализа национального корпуса путем различных запросов. В поисковик входят следующие приложения: приложение сбора данных; приложение индексирования и поиска; приложение структурирования исходных данных. Поисковик был создан на языке Java и выполнен в Desktop варианте, что позволяет установить его на компьютере лингвиста-исследователя. Достоинствами созданного поисковика являются: бесплатная лицензия и относительное быстродействие. Недостатком является: необходимость установки дополнительных программ.

Ключевые слова: поисковик (search engine), текстовый корпус (text corpora), разметка текста, запрос (query), индексирование (indexing).

Petrov Alexander Vasilyevich,
student,
«Chuvash State University
named after I.N. Ulyanov»
Cheboxary

DEVELOPMENT OF THE LEXICAL SEARCH ENGINE ARCHITECTURE FOR THE CHUVASH LANGUAGE NATIONAL CORPUS IN THE JAVA LANGUAGE

ANNOTATION. The article describes in detail the architecture and functional of the search engine for the Chuvash language national corpus, adjusted for all requirements and specificities. The search engine is a specialized

program for analyzing the national corpus by means of various queries. The search engine includes the following applications: data collection application; indexing and retrieval application; primary data structuring application. The search engine was programmed in the Java language and was made in a Desktop variation, which allows to install it on a linguist researcher's computer. The benefits of the created search engine are its free license and relatively high operating speed. Its shortcoming is a necessity of installing additional programs.

Keywords: search engine, atrial fibrillation, an international normalized ratio (INR)

В настоящее время для сохранения текстового и лексического богатства национальных языков создаются национальные корпуса, которые являются огромными структурированными электронными хранилищами текстов с возможностью быстрого поиска на нескольких уровнях языка: морфемном, морфологическом, синтаксическом, текстовом и семантическом.

Быстрый поиск в таких корпусах осуществляется с помощью поисковиков.

Поисковик представляет собой специализированную программу для анализа национального корпуса путем различных запросов.

Основной задачей поисковика является предоставление исследователям возможностей по сбору художественных текстов в автоматизированное информационное хранилище, их исследование в различных плоскостях и использование текстов/результатов их анализа в своих работах.

Соответственно, возникают следующие достаточно формальные задачи:

(1) Задача сбора и индексации художественных текстов;

(2) Задача поиска художественных текстов;

(3) Задача анализа найденных художественных текстов;

(4) Задача визуализации найденных художественных текстов.

Нами предполагается, что система сбора и структурирования художественных текстов нужна для последующего ретроспективного поиска заданных пользователем слов в предложениях из художественных произведений. В основе автоматизированной системы лежат методы анализа текстов. Система загружает данные из художественных произведений. Загрузив данные, система структурирует их и применяет к ним методы анализа текстов, а затем выдает результаты анализа пользователю системы.

С учетом обозначенных принципов нами была разработана архитектура поисковика для национального корпуса чувашского языка и реализован сам поисковик.[3]

Поисковик был создан на языке Java и в настоящее время функционирует в Desktop варианте.

Поисковик представляет собой систему, в которую входят следующие приложения:

1) приложение сбора данных;

2) приложение индексирования и поиска;

3) приложение структурирования исходных данных.

Перейдем к описанию разработанных компонент поисковика.

1. Приложение сбора данных

Цель сбора данных – собрать тексты художественных произведений, которые пользователь собирается исследовать.

Общая схема предлагаемого нами приложения сбора данных состоит из двух частей: клиентской и серверной. Клиентская часть представляет собой веб-интерфейс, при помощи которого пользователь может: загрузить файл художественного произведения на сервер вместе со всей со-

путствующей информацией (автор произведения, название произведения и т.п.); просмотреть все загруженные на сервер художественные произведения; удалить художественное произведение с сервера. Серверная часть обрабатывает поступающие от пользователя запросы и выдает ему результаты запросов.

2. Приложение индексирования и поиска

Итак, как было сказано ранее, для работы с большими объёмами данных необходимы большие вычислительные мощности и объёмы дискового пространства. Массив оцифрованных художественных текстов очень велик, и поиск требуемых предложений без предварительной обработки текстов требует больших вычислительных ресурсов и длительного времени. Поэтому в системе используется индексирование предложений художественных текстов. Суть индексирования заключается в возможности добавлять, удалять или обновлять документы в хранилище данных (1 документ = 1 предложение), впоследствии использующегося для полнотекстового поиска информации. [1]

3. Приложение структурирования исходных данных

Логические сущности, которые будут фигурировать в базе данных: автор, художественный текст, предложение. Автор – персона, создающая художественные тексты. Каждый художественный текст может быть привязан к нескольким авторам, а также каждый автор может быть привязан к разным художественным текстам (связь многие-ко-многим). Художественные тексты состоят из предложений. Каждое предложение связано с единственным художественным текстом (связь многие-ко-одному). [2]

Таким образом на данном этапе разработана лексическая база языка, которая может пополняться художественными произведениями и автоматически структурировать их.

СПИСОК ЛИТЕРАТУРЫ

1. Желтов П.В. Лингвистические процессоры, формальные модели и методы: теория и практика / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 208 с.

2. Желтов П.В. Формальные методы в сравнительно-сопоставительном языкознании / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 252 с.

3. Желтов П.В. Лингвистические процессоры в системах искусственного интеллекта / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2007. – 100 с.