Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники
Кафедра инженерной психологии и эргономики

На правах рукописи

УДК 004.5:004.42

Заяц
Дмитрий Викторович

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

Автореферат диссертации на соискание академической степени
магистра технических наук

1 - 23 80 08 Психология труда, инженерная психология, эргономика

Магистрант Д.В. Заяц

Научный руководитель
Л.Т. Ткачева, кандидат
технических наук, доцент

Заведующий кафедрой ИПиЭ
К.Д. Яшин, кандидат
технических наук, доцент

Нормоконтролёр
Т.В. Гордейчук,
Ассистент кафедры ИПиЭ,
магистр технических наук

Минск 2016

ВВЕДЕНИЕ

С момента появления компьютеров человек пытался возложить на него свои обязанности, максимально использовать его для выполнения ежедневных рутинных операций. За многие годы, с появлением множества разнообразного программного обеспечения, компьютер, за счет точности и детерминированности вычислений, стал достойной заменой человека при проведении многих математических и физических расчетов, выполнению различных шаблонных действий, таких как форматирование документов, построение чертежей и многих других операций, требующих единого подхода. Однако, все еще остается множество операций, для которых компьютеру по-прежнему не хватает человеческого восприятия, опыта, который каждый человек получает с первых дней своей жизни. Такими обыденными для человека действиями, пока не до конца подвластными компьютеру, можно считать распознавание изображений, обработку естественных языков (тех языков, на которых говорят люди) и прочие действия, для которых важен накопленный людьми опыт, а также немного интуиции. Компьютеру порой не достает единого механизма принятия решений, которым, наверняка, пользуется человек. Скорее всего, именно накопление этого опыта, а также принятие решений и являются «китами», на которых может быть создан искусственный интеллект.

Объем информации, размещенной во Всемирной сети, стремительно растет, однако, ее качество и безопасность могут вызывать опасения. Если во времена, когда Интернет делал еще первые шаги, и главными создателями содержимого сайтов являлись их владельцы, злоумышленника можно было относительно просто вычислить, то с популяризацией методики проектирования систем Web 2.0, когда доступ к созданию контента получило множество анонимных пользователей, остро встал вопрос фильтрации нежелательной информации.

В настоящее время, несмотря на растущую долю мультимедийного контента во всемирной сети, по-прежнему множество информации представлено в текстовом формате. Сегодня это наиболее простой и доступный для рядового пользователя персонального компьютера способ создания цифровой информации. Тем не менее, этот формат представления информации является самым «компактным» по соотношению «размер/полезная информация», что делает его наиболее удобным для автоматической обработки.

Темой данной работы назначено создание программы, которая могла бы обрабатывать тексты на русском языке, собирать статистику использования в них различных слов и фраз, определять статистические зависимости между текстами одной тематики, а также, на основании имеющихся данных, делать предположения о тематике ранее неизвестных программе текстов.

Программа впоследствии может быть адаптирована для использования в качестве фильтров для нежелательной информации в сети. Сохраняя основную концепцию определения тематики, необходимым изменением при интеграции с конкретным информационным ресурсом должно быть создание моста, по которому информация будет получаться из ресурса для дальнейшей передачи в модуль анализатора.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объект исследования – компьютерная лингвистика: методы решения задач компьютерной морфологии, обработки естественного языка на уровне предложения и текста.

Предмет исследования – программа, позволяющая анализировать русскоязычный текст: выделять слова, словосочетания и фразы, накапливать статистику использования слов и фраз в различных текстах. Кроме того, предметом исследования является поиск статистических зависимостей между текстами подобной тематики, с целью распознавания тематики произвольных текстов.

Цель диссертационной работы – разработка приложения для статистического анализа русского текста.

Способами достижения цели выступают конкретные исследовательские задачи:

- проанализировать техническую литературу и современные аналоги;
- создать структуру базы данных;
- разработать алгоритмы программы, осуществляющие поиск статистических зависимостей между текстами подобной тематики;
- разработать пользовательский интерфейс;
- разработать клиентское приложения для реализации необходимых задач.

Магистерская диссертация является завершенной, поставленная задача решена в полной мере, присутствует возможность дальнейшего развития системы и увеличение её функционала.

Результаты работы доложены на 51-й научно-технической конференции студентов, магистрантов, аспирантов БГУИР в 2015 году.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе магистерской работы проводится обзор предметной области, теоретический анализ технической литературы, обзор эргономических практик и проблем современности и постановка задач для разработки. Основным вопросом, затронутыми в исследовании, является оптимальный способ построения комплекса программ, объединяющего в себе как средства для компьютерной лингвистики на нескольких уровнях (слово, предложение, текст), так и средств прикладного уровня — как пользующихся результатами работы лингвистических программ, так и предоставляющих новые данные для уточнения работы таких программ. Немаловажным является вопрос производительности лингвистических программ: важно найти оптимальное соотношение трудоемкости анализа текста с качеством получаемого результата. Поскольку целью работы комплекса не является полная замена человеческого труда, а ее заключения носят рекомендательный характер, качеством получаемого результата можно пренебречь с целью получения актуальных результатов в режиме онлайн.

При постанови задач для разработки программного комплекса определена главная цель, входные и выходные параметры, а также основные требования к программному комплексу.

Проанализировав доступные источники, а также существующие готовые решения было выяснено, что для создания программы для статистического анализа текстов, а впоследствии и интегрируемого в сайт модуля, требуется решить следующие задачи:

- разделение текста на слова и предложения;
- определение начальной формы слова;
- вычисление статистических зависимостей между текстами;
- вывод предположений о новых текстах на основании полученных данных.

Важным фактором является определение методов компьютерной лингвистики, которые будут использоваться в программе – основанные на определенных правилах, либо на основании специальной обученной с помощью языковых корпусов модели. Следующей задачей является сбор статистики по исследуемым текстам. Данные для этой задачи являются специфичными для каждого пользователя, таким образом, в программу должен быть заложен алгоритм, который реализует сбор этой информации.

Во второй главе производится разбор основных технологий и методов, использующихся для разделения текстов на слова и предложения, определения начальной формы слов, вычисления статистических зависимостей таких, как метод опорных векторов, латентно-семантический анализ и Стеммер Портера.

В третьей главе производится обоснование выбора средств разработки, описание архитектуры и логических модулей приложения. Также в этой главе описывается разработка и тестирование программного средства. Во время анализа современных технологий и методов были выявлены как их достоинства, так и недостатки. Несмотря на ряд некоторых минусов был выбран язык разработки Java. В качестве архитектуры программного средства была выбрана многоуровневая модель. Архитектура программного продукта представляет из себя набор взаимосвязанных компонентов и модулей, базы данных.

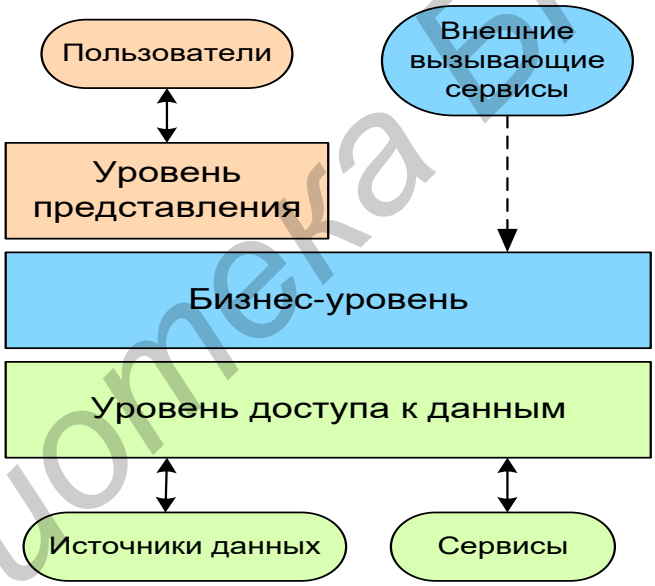

Рисунок 1 – Уровни бизнес-приложения

Многоуровневая архитектура (рисунок 1) сосредоточена на иерархическом распределении отдельных частей системы при помощи эффективного разделения отношений. Каждая часть соотносится с определенным уровнем, для каждого уровня заданы выполняемые им функции, уровни выстроены в стековую структуру (то есть находятся один поверх другого). Например, типичная архитектура для веб-приложений включает уровень представления, уровень бизнес-логики и уровень доступа к данным.

ЗАКЛЮЧЕНИЕ

В ходе работы были исследованы задачи и проблемы компьютерной лингвистики, изучены различные подходы к решению задач компьютерной морфологии русского языка и статистического сравнения текстов. На основе данного анализа были выбраны необходимые алгоритмы и технологии, с помощью которых была спроектирована и разработана программа для статистического анализа текстов.

Программа разработана с использованием средств платформы Java и языка Python. Эти технологии позволяют использовать программу в любой операционной системе, для которой существуют реализация JVM и интерпретатор Python. Программа может использоваться как модуль другой системы (например, спам-фильтр сайта) или как независимое приложение с собственным пользовательским интерфейсом.

Достоинствами системы являются:

- простота и удобство использования;
- платформонезависимость;
- возможность использования и в качестве библиотеки или модуля системы, и как самостоятельное приложение (например, для анализа текстов при проведении социологических исследований);
- свободное распространение пробной версии программы для академических нужд.

Недостатками программы являются:

- недостаточная оптимизация производительности программы;
- как следствие первого недостатка, не максимальное возможное качество решения задач компьютерной морфологии в результате выбора более простых алгоритмов.

Тем не менее, программа обладает потенциалом к развитию за счет применения более точных алгоритмов компьютерной лингвистики, а также использования некоторых алгоритмических эвристик при реализации этих алгоритмов.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] Заяц, Д.В. Статистический анализ текстов / Д.В. Заяц // 51-я научная конференция аспирантов, магистрантов и студентов. – Минск, 2015.