МИНОБРНАУКИ РОССИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет *романо-германской филологии*

Кафедра *французской филологии*

Направление *«Лингвистика»*

Профиль *«Теория и методика преподавания иностранных языков и культур»*

Проект на тему:

«Информационные технологии в обработке текстов. Автоматическая обработка текстов. Автоматическое распознавание текста. Автоматическое аннотирование и реферирование текста. Автоматический анализ и синтез текста. Формулировка задачи автоматического реферирования и аннотирования текста. Системы автоматического реферирования, аннотирования и членения текста.»

Выполнила: студентка 1 курса 1 группы Мурашко Вероника Васильевна

Руководитель: Донина Ольга Валерьевна

Г. Воронеж

2018 год

Оглавление

Введение

Практически каждый встречается с необходимостью подготовки тех или иных текстовых документов (справок, служебных записок, отчетов, статей, писем, рекламных материалов и т.п.). Текстовые документы можно подготовить и без компьютера, например, на пишущей машинке. Однако с появлением компьютеров готовить документы стало значительно проще и удобнее. Принципиальное отличие создания и редактирования текста на компьютере от традиционных технологий состоит в том, что технически не представляет труда устранить ошибку в тексте немедленно, а также и впоследствии. Это создает большое психологическое преимущество пользователю, избавляя его от опасений допустить ошибки, исправление которых при старых технологиях требует определенных усилий и времени.

Автоматизированные информационные технологии (АИТ) обработки текстовой информации – одна из базовых информационных технологий в современном мире. От правильности составления документов и их оформления во многом зависит деятельность любой современной организации. Внешний облик и оформление наряду с содержанием документов могут многое рассказать об организации. Они являются своего рода показателем информационной культуры компании, составляющими ее имиджа. Поэтому чрезвычайно важно в полной мере владеть технологией обработки текстовой информации. Умение правильно оформить документ в соответствии с предъявляемыми к нему требованиями является одним из критериев оценки специалиста в области информационных технологий.

Общее название программных средств, предназначенных для создания, редактирования и форматирования простых и комплексных текстовых документов — *текстовые процессоры*. В настоящее время в России наибольшее распространение имеет текстовый процессор MS Word1. Это связано, прежде всего, с тем, что его создатели относительно
давно предусмотрели локализацию программы в России путем включения
в нее средств поддержки работы с документами, исполненными на русском языке.

Основная часть

1.Автоматическая обработка текстов.

Автоматическая обработка текста (АОТ) —

преобразование текста на искусственном или естественном языке с помощью ЭВМ. Прикладные системы и теория АОТ начали создаваться в конце 50-х гг. 20 в. (США, СССР, Франция, ФРГ и др.) и развивались в нескольких различных приложениях: в системном программировании, издательском деле и в вычислительной лингвистике. В системном программировании, предметом которого является создание программного обеспечения функционирования ЭВМ и работы пользователей, развивались инструментальные средства разработки программ, т. е. текстов на алгоритмических языках. В издательском деле АОТ— одно из направлений автоматизации редакционно-издательских процессов. В этих областях термин «АОТ» употребляется, как правило, в относительно узком смысле как преобразование формы. В вычислительной лингвистике, предметом которой является автоматический лингвистический анализ и синтез текста, а также лингвистические аспекты общения с ЭВМ на естественном языке, термин «АОТ» понимается в более широком смысле, охватывающем и процедуры анализа содержания и синтеза (по заданному содержанию понятного человеку) текста.

В зависимости от целей различают несколько видов АОТ. Преобразование текстов при автоматизированном редактировании заключается во внесении в текст, находящийся в памяти ЭВМ, исправлений и дополнений; форматирование текста заключается в выделении заголовков, формировании строк и страниц

нужного формата, выделении и оформлении разделов и подразделов текста для его воспроизведения на устройствах печати ЭВМ.
В процессе автоматического набора и вёрстки текст, введённый в ЭВМ, преобразуется в представление (код), воспроизводимое полиграфическим оборудованием (например, фотонаборным автоматом). При лексикографической обработке текст преобразуется в лексикографическое представление, в котором каждому словоупотреблению соответствует определённая информация в формируемом к этому тексту словаре.
В автоматическом лингвистическом анализе текст последовательно преобразуется в его лексемно-морфологическое, синтаксическое и семантическое представления.
В процессе автоматического синтеза производятся обратные преобразования: от семантического представления через синтаксическое и лексемно-морфологическое к собственно текстовому.

2. Автоматическое распознавание текста.

После обработки документа сканером получается графическое изображение документа (графический образ). Но графический образ еще не является текстовым документом. Человеку достаточно взглянуть на лист бумаги с текстом, чтобы понять, что на нем написано. С точки зрения компьютера, документ после сканирования превращается в набор разноцветных точек, а вовсе не в текстовый документ.

Проблема распознавания текста в составе точечного графического изображения является весьма сложной. Подобные задачи решаются с помощью специальных программных средств, называемых *средствами распознавания образов*. Реальный технический прорыв в этой области произошел лишь в последние годы. До этого распознавание текста было возможно только путем сравнения обнаруженных конфигураций точек со стандартным образцом (эталоном, хранящимся в памяти компьютера). Подобные системы назывались OCR (Optical Character Recognition – оптическое распознавание символов) и опирались на специально разработанные шрифты, облегчавшие такой подход.

Программа FineReader выпускается отечественной компанией ABBYY Software. Эта программа предназначена для распознавания текстов на русском, английском, немецком, украинском, французском и многих других языках, а также для распознавания смешанных двуязычных текстов.

Программа имеет ряд удобных возможностей. Она позволяет объединять сканирование и распознавание в одну операцию, работать с пакетами документов (или с многостраничными документами) и с бланками. Она позволяет редактировать распознанный текст и проверять его орфографию.

Панель *Scan&Read* содержит кнопки, соответствующие всем этапам превращения бумажного документа в электронный текст. Первая кнопка позволяет выполнить такое преобразование в единой операции. Остальные кнопки соответствуют отдельным этапам работы и содержат раскрывающиеся меню, служащие для управления соответствующей операцией.

Панель *Распознавание* позволяет указать язык документа и вид шрифта. Последнее требуется делать только в тех случаях, когда документ имеет недостаточное качество печати.

Панель *Инструменты* используют при работе с исходным изображением. В частности, она позволяет управлять сегментацией документа. С помощью элементов управления этой панели задают последовательность фрагментов текста в итоговом документе.

Элементы управления панели *Форматирование* используют для изменения представления готового текста или при его редактировании.

3. Автоматическое аннотирование и реферирование текста.

Рефератом называется связный текст, который кратко выражает не только центральную тему или предмет какого-либо документа, но и цель, применяемые методы, основные результаты описанного исследования или разработки. Рефераты обычно составляют к научно-техническим документам (научным книгам, статьям, патентам на изобретение и т. п.) Реферат акцентирует внимание читателя на новых сведениях и определяет целесообразность его обращения к исходному документу. Он помогает человеку ориентироваться в

информационных потоках, оперативно отбирать для себя наиболее ценную и полезную информацию. Процесс составления реферата называется *реферированием*.

Аннотацией называют краткое изложение содержания документа, дающее общее представление о его теме. Таким образом, если реферат в краткой форме знакомит читателя с сутью излагаемого в документе содержания (фактами, методикой, экспериментами и т. п.), то аннотация выполняет лишь сигнальную функцию, сообщая о том, что опубликована статья или книга на определенную тему. Процесс составления аннотации называется *аннотированием*.

Рефераты и аннотации представляют собой вторичные документы (первичные, или исходные, документы — это книги, статьи, патенты и т. п.) В каждом вторичном документе можно выделить два компонента информации: содержательный и документо-графический. Первый компонент содержит информацию первоисточника (о чем книга, статья). Второй компонент — это сведения о самом первичном документе (тип документа: книга, статья и т. п.; вид: печатный, рукописный; год издания; место издания и т. д.).

Реферирование и аннотирование текста являются сложными видами интеллектуальной деятельности. Составление человеком рефератов или аннотаций занимает много времени. Это приводит к тому, что до ученых, педагогов, инженеров и других специалистов новейшая информация (особенно зарубежная) доходит очень медленно, что, в свою очередь, ведет к повторению в разных странах и в пределах одной страны одних и тех же исследований, более позднему применению новейших методик, технологий, процессов. Чтобы как-то избежать этого, для составления рефератов и аннотаций применяют современные компьютеры.

Составление реферата или аннотации текста с помощью компьютера называется *автоматическим реферированием* или *аннотированием.*

3.1.Три этапа выполнения работы по составлению реферата или аннотации.

1) подготовительный — референт определяет тематическую направленность текста и пытается понять и осмыслить документ в целом;

2) аналитический — референт делит текст на некоторые фрагменты (абзацы, аспекты и т. п.). Каждый фрагмент внимательно изучается, в нем выделяют основные смысловые единицы (предложения, словосочетания, слова);

3) этап непосредственного построения реферата или аннотации — выделенные ранее смысловые единицы (их комбинации или преобразования) располагаются в единый вторичный текст в соответствии с планом реферата или аннотации.

В качестве основных смысловых единиц, выделяемых из исходного текста на 2-м этапе, могут выступать:

- целые ключевые предложения;
- ключевые словосочетания и слова.

Ключевое (опорное) слово — это термин, относящийся к основному содержанию текста и повторяющийся в нем несколько раз (с учетом всех возможных синонимов).
Ключевое словосочетание — это сочетание слов, среди которых есть одно или несколько ключевых.
Ключевым предложением считается предложение, содержащее два и более ключевых слова или ключевых словосочетания.

Создаваемый на 3-м этапе реферат или аннотация содержат выделенные ранее смысловые единицы.

3.2.Смысловые единицы реферата и аннотации.

В качестве смысловых единиц реферата могут выступать:

1) полные (без изменения) ключевые предложения исходного текста;
2) перефразированные ключевые предложения исходного текста;
3) предложения, составленные из ключевых слов или словосочетаний исходного текста с помощью специальных связующих элементов;
4) предложения, обобщающие несколько предложений исходного текста (не обязательно ключевых).

Смысловыми единицами аннотации могут быть:

1) ключевые слова или словосочетания исходного текста с предшествующими им специальными фразами — реляторами типа: «В статье рассматриваются следующие вопросы:...», «Книга посвящена следующим проблемам: ...» и т. п.;

2) специальные предложения, содержащие оценочные элементы: «Рассматривается важная проблема...», «Статья посвящена актуальной теме...» и т. д.;

3) специальные предложения, содержащие клише, т. е. специализированные словесные штампы, фиксирующие внимание читателя на определенных аспектах содержания: «Недостаток... заключается», «Цель публикации...», «Ставится задача...», «Делается попытка...» и т. д.

4.Автоматический анализ текста.

Автоматический анализ текста представляет собой операцию, которая из заданного текста на естественном языке извлекает грамматическую и семантическую информацию, содержащуюся в тексте. Автоматический анализ выполняется по некоторому алгоритму в соответствии с заранее разработанным описанием данного языка. Обратная операция называется автоматическим синтезом текста.

Автоматический анализ является одним из важнейших этапов в различных видах автоматической обработки текстов:

· автоматического реферирования;

· автоматического перевода;

· информационного поиска.

Автоматический анализ не стоит путать с автоматическим исследованием текстов, в котором практически полностью отсутствуют данные о языке обрабатываемого текста, и обработка текста осуществляется алгоритмом с целью создания описания языка. В алгоритмах автоматического анализа, как правило, имеются сведения о языке (его «грамматика») и сведения о самом процессе анализа («механизм», т.е. алгоритм автоматического анализа).

Любая современная система анализа текста, в том числе поисковые машины, осуществляющие поиск документов в сети Интернет, содержит те или иные модули автоматического лингвистического анализа. Необходимыми этапами лингвистического анализа практически в любой современной системе являются:

· токенизация (разбиение на орфографические слова и выделение границ предложений);

· морфологический анализ (разбор слова как части речи).

5. Формулировка задачи автоматического реферирования и аннотирования текста.

Отметим следующие новые задачи, связанные с компьютерным реферированием:

1. Создание одноязычных рефератов из источников на разных языках.

2. Построение рефератов по гибридным источникам, включающим как текстовые, так и числовые данные в разных формах (таблицы, диаграммы, графики и т. д.).

3. Создание рефератов на основе массивов документов. Например, по строение единого реферата по сборнику тезисов докладов научной конференции. Одна из областей применения подобных средств —

формирование новостных сообщений по газетным источникам.

4. Растущий объем мультимедийной информации обусловливает актуальность разработки средств ее автоматического реферирования. Методы извлечения семантики из мультимедийной информации находятся на начальных стадиях развития.

Средства автоматического аннотирования в целом аналогичны средствам автоматического реферирования. Однако требования к сжатию текста для них, как правило, на порядок более жесткие.

6. Системы автоматического реферирования, аннотирования и членения текста.

К традиционным системам автоматического реферирования и аннотирования, реализующим поверхностные методы, можно отнести:

- Microsoft Word (функция автоматического реферирования);
- ОРФО 5.0 (разработчик — компания «Информатик»), включающую функцию автоматического аннотирования русских текстов;
- «Либретто» (разработчик — компания «МедиаЛингва»), обеспечивающую автоматическое реферирование и аннотирование русских и англий ских текстов (система встраивается в Word);
- пакет «МедиаЛингва Аннотатор SDK 1.0», служащий инструментарием для реализации функций автоматического реферирования и аннотиро вания в прикладных ИАС;
- поисковую систему «Следопыт», включающую средства автоматического реферирования и аннотирования документов.
-
- поисковую машину «Золотой Ключик» компании Textar, обеспечивающую составление рефератов и аннотаций;
- Intelligent Text Miner (IBM);
- Oracle Context;
- программные компоненты для разработки систем управления знаниями Inxight Summarizer фирмы Inxight Software, Inc.

Практическая работа

Задание №1

Порядок выполнения:

1) Ответьте на вопросы:

Из каких структурных элементов состоит интерфейс текстового редактора?
Расскажите в чем суть операций копирования, перемещения, и удаления фрагмента текста?
Для каких целей проводится выделение фрагментов текста?
В чем сущность режима нахождения и замены?
Как производится проверка правописания слов?
В чем сущность режима проверки синтаксиса и стиля?
Как и для какой цели производится форматирование документа?
В чем сходство и различие текстового редактора и издательской системы?
Каким образом задаются параметры и нумерация страниц?

2) Найдите учебник по ИТ Григорьева Е.С. и выполните Лабораторную работу №1 «Форматирование символов и текста». Документ в формате PDF выложите на GitHub.
3) Составьте принципиальный алгоритм решения задачи автоматического реферирования и аннотирования текста.

Закрепление материала

Задание №1

Вопросы:

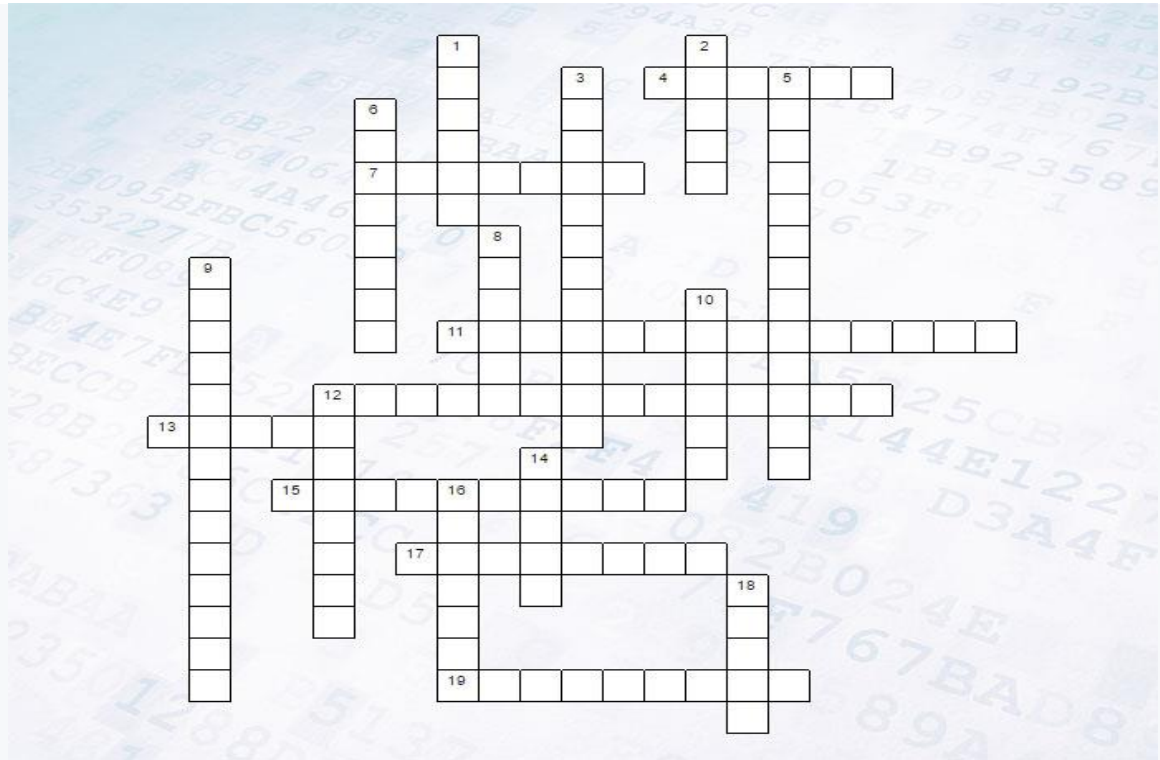
По горизонтали:

4. Форматирование, применяемое к произвольным символьным фрагментам (отдельным символам, словам, строкам, предложениям) и абзацам. 7. Ориентация листа бумаги, при которой высота листа больше его ширины. 11. Этап создания текстового документа, на котором его просматривают, исправляют обнаруженные ошибки и вносят необходимые изменения. 12. Процесс оформления текста. 13. Промежуточное хранилище данных, предоставляемое программным обеспечением и предназначенное для переноса или копирования

между приложениями через операции Вырезать, Копировать, Вставить. 15. Часть текстового документа, составленная из названий разделов определённых уровней. 17. Форматирование, позволяющее быстро изменить стиль одинаковых структурных элементов во всем документе. 19. Ориентация листа бумаги, при которой ширина листа больше его высоты.

По вертикали:

1. Наклонное начертание символов. 2. Выполненные в едином стиле изображения символов, используемых для письма. 3. Расположение абзаца относительно боковых границ страницы. 5. Интервал, определяющий расстояние между соседними строками внутри абзаца. 6. Текстовая информация, представленная на бумажном, электронном или ином материальном носителе. 8. Режим работы текстового редактора, при котором символ, стоящий за курсором, заменяется символом, вводимым с клавиатуры. 9. Список, элемент которого сам является списком. 10. Минимальная графическая единица текста. 12. Произвольное количество следующих один за другим символов текста. 14. Размер шрифта. 16. Режим работы текстового редактора, при котором существующий текст сдвигается вправо, освобождая место вводимому тексту. 18. Часть документа между двумя соседними непечатаемыми управляющими символами, получаемыми при нажатии клавиши Ente



Ответы на кроссворд Обработка текстовой информации:
По горизонтали: 4. Прямое. 7. Книжная. 11. Редактирование. 12. Форматирование. 13. Буфер. 15. Оглавление. 17. Стилевое. 19. Альбомная.
По вертикали: 1. Курсив. 2. Шрифт. 3. Выравнивание. 5. Междустрочный. 6. Документ. 8. Замена. 9. Многоуровневый. 10. Символ. 12. Фрагмент. 14. Кегль. 16. Вставка. 18. Абзац.

Задание №2

Сопоставьте названия терминов с и определением.

1	Реферат	А	операция, которая из заданного текста на естественном языке извлекает грамматическую и семантическую информацию, содержащуюся в тексте.
2	Аннотация	Б	смысловые единицы (их комбинации или преобразования) располагаются в единый вторичный текст в соответствии с планом реферата или аннотации.
3	Автоматический анализ текста	В	связный текст, который кратко выражает не только центральную тему или предмет какого-либо документа, но и цель, применяемые методы, основные результаты описанного исследования или разработки.
4	этап непосредственного построения реферата или аннотации	Г	краткое изложение содержания документа, дающее общее представление о его теме.

Ответы:

1-В

2-Г

3-А

4-Б

Задание №3

Перечислите смысловые единицы реферата:

Ответ:

1) полные (без изменения) ключевые предложения исходного текста;
2) перефразированные ключевые предложения исходного текста;
3) предложения, составленные из ключевых слов или словосочетаний исходного текста с помощью специальных связующих элементов;
4) предложения, обобщающие несколько предложений исходного текста (не обязательно ключевых).

Задание№4

Назовите системы автоматического реферирования, аннотирования и членения текста.

Ответ:
Microsoft Word (функция автоматического реферирования);
Intelligent Text Miner (IBM);
Oracle Context и др.

Список используемых источников

1) Агафонова И.В., Дмитриева О.В. Эволюция шрифтов замены: идеи и механизмы. Часть 1. - СПб.: Изд-во ЦПО "Информатизация образования", 2006, N5, С. 9-15.

2) Современный редактор текстов. Под редакцией А.Н.Лучника Москва 2004 г.

3) https://revolution.allbest.ru/programming/00768621_0.html

4) https://knowledge.allbest.ru/programming/3c0b65635a2ad78a5d43a89421216c26_0.html

5) https://www.bestreferat.ru/referat-311722.html

6) http://journalpro.ru/articles/avtomaticheskiy-analiz-tekstov-sintaksicheskiy-i-semanticheskiy-analiz/

7) https://vuzlit.ru/1013608/avtomaticheskoe_raspoznavanie_tekstov