

*ТЕХНОЛОГИЯ АНАЛИЗА
ТЕКСТА
И ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ
СЛОВ*

Студент группы 320602
Котиков Е.В.

ЦЕЛЬ РАБОТЫ

- практическое освоение технологии анализа текста;
- извлечение ключевых слов;
- профессионального поиска информации.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

- *Знание общих принципов функционирования поисковых средств и умение грамотно составить запрос поисковой машине необходимые, но недостаточные условия успешного поиска требуемой информации.*
- *Выбор ключевых слов в данном случае может осуществить специалист узкого профиля, но труд его дорог и малопроизводителен, или специальные программные средства, основанные на применении законов Зипфа.*

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

- Джордж Зипф установил, что все тексты подчиняются общим закономерностям, и сформулировал в 1946—49 гг. несколько законов, которые нашли применение в технологии поиска информации.
- Первый закон Зипфа утверждает, что произведение частоты встречи слова в тексте (или вероятности встречи слова по Зипфу) на его ранг есть величина приблизительно постоянная для любых текстов определенного языка, т.е. имеет место $C = f \cdot R \sim \text{const.}$

График зависимости частоты слова f от его ранга R .

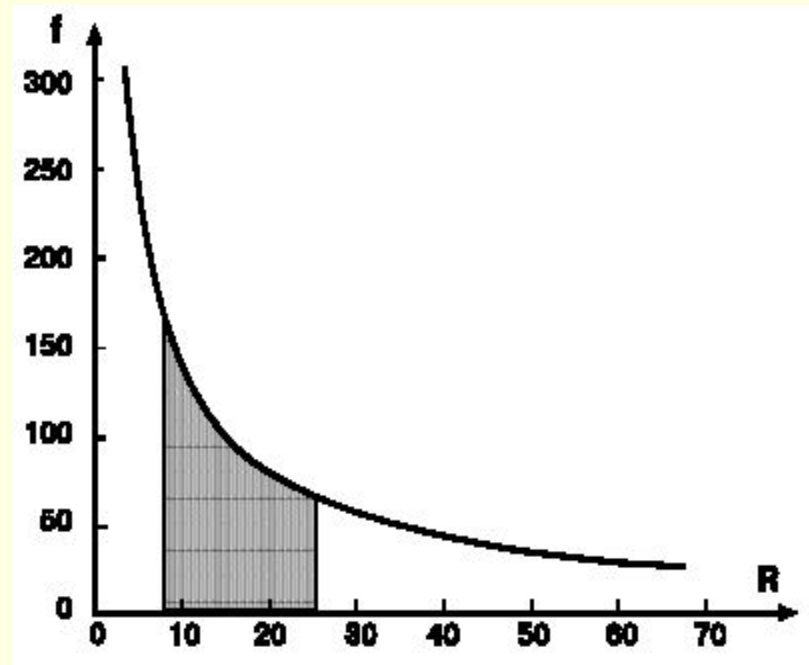


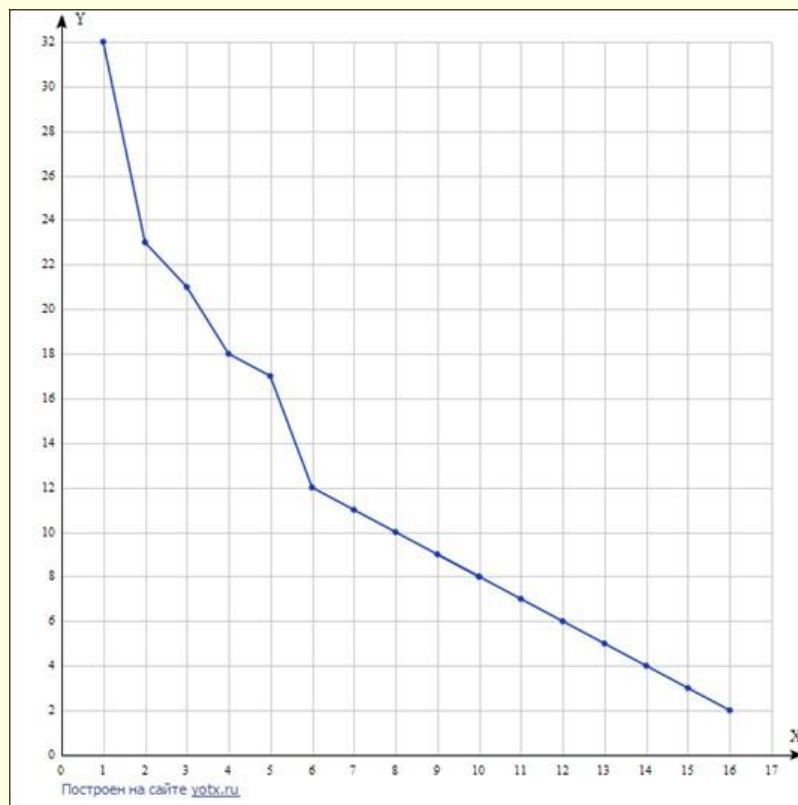
ТАБЛИЦА КЛЮЧЕВЫХ СЛОВ (ручной поиск)

Список слов текста-источника, отсортированный по убыванию их частот, представлен в следующей таблице

Ключевые слова	f	R
ERP	32	1
клиент	23	2
знания	21	3
бизнес-процесс	18	4
компания	17	5
информация	12	6
ERP-система	11	7
системы	11	7
управление	11	7
процессы	10	8

График зависимости частоты вхождения слова от ранга

В данном случае целесообразно выбрать диапазон значений ранга слов, равный первым восьми рангам (10 слов).



ФОРМУЛИРОВКА ПОИСКОВОГО ЗАПРОСА (ручной поиск)

Запрос будет состоять из слов с наивысшим рангом, разделенных логическим оператором «ИЛИ» (or). При этом порядок следования сохраним.

The screenshot shows a Google search interface with the following elements:

- Address Bar:** https://www.google.by/?gws_rd=ssl#q=ERP+or+%D0%BA%D0%BB%D0%B8%D0%B5%D0%BD%D1%82+or+%D0%B7%D0%BD%D0%B0%D0%BD%D0%B8%D1%8F
- Search Bar:** Contains the query "ERP or клиент or знания or бизнес-процесс or компания or информ".
- Navigation:** Links for "Поиск", "Новости", "Видео", "Картинки", "Карты", "Ещё", and "Инструменты поиска".
- Results:**
 - Count: "Результатов: примерно 760 (1,38 сек.)"
 - Warning: "or" и все последующие слова были проигнорированы, так как количество слов в запросе не может превышать 32."
 - Result 1: **Эволюция технологий ERP-систем**
capri.urfu.ru/intelligentE/vershiny.htm
В архитектуре ERP II много технических и функциональных уровней, но одна из ... отрасли и повышающие ценность компании для клиентов и владельцев за счет ... бэк-офисное ERP-приложение из внутренней бизнес-системы в ... SCM (Supply Chain Management), а также системы управления знаниями ...
 - Result 2: **8.1.1.4 SAP ERP HCM - Информационные технологии ...**
khp-iip.mipk.kharkiv.edu/library/itob/itob08.html
8.1.1 SAP ERP - управление ресурсами предприятия. 8.1.1.1 ... 8.1.2 SAP SCM - управление логистической сетью ... 8.2.2 Oracle CPM - эффективное управление бизнесом ... SAP Business Suite предоставляет решения, поддерживающие ... Создание стоимости и поддержка ключевых процессов оперативной ...
 - Result 3: **PDF Информационные технологии в бизнесе Часть 1.pdf**
elib.bsui.by/.../Информационные%20технологии%20в%20бизнесе%20...
автор: ЛВ Белецкая - 2012 - Похожие статьи
и ИС поддержки исполнения: классификация и различия ... Информационные технологии в системе управления ... и принятия решений в изменившихся условиях; ... возрастанием роли информации и знаний в жизни общества; ... отдельным видам бизнес-процессов, созданием информационных баз ...

Программы-экстракторы

- RCO Fact Extractor – это интеллектуальная программа для высокоточного избирательного анализа информации.
- TextAnalyst – персональная система автоматического анализа текста, разработан а качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов.

ИСПОЛЬЗОВАНИЕ TextAnalyst

Персональная система автоматического анализа текста **TextAnalyst** предназначена для анализа содержания текстов, смыслового поиска информации и формирования электронных архивов. TextAnalyst предоставляет пользователю следующие возможности:

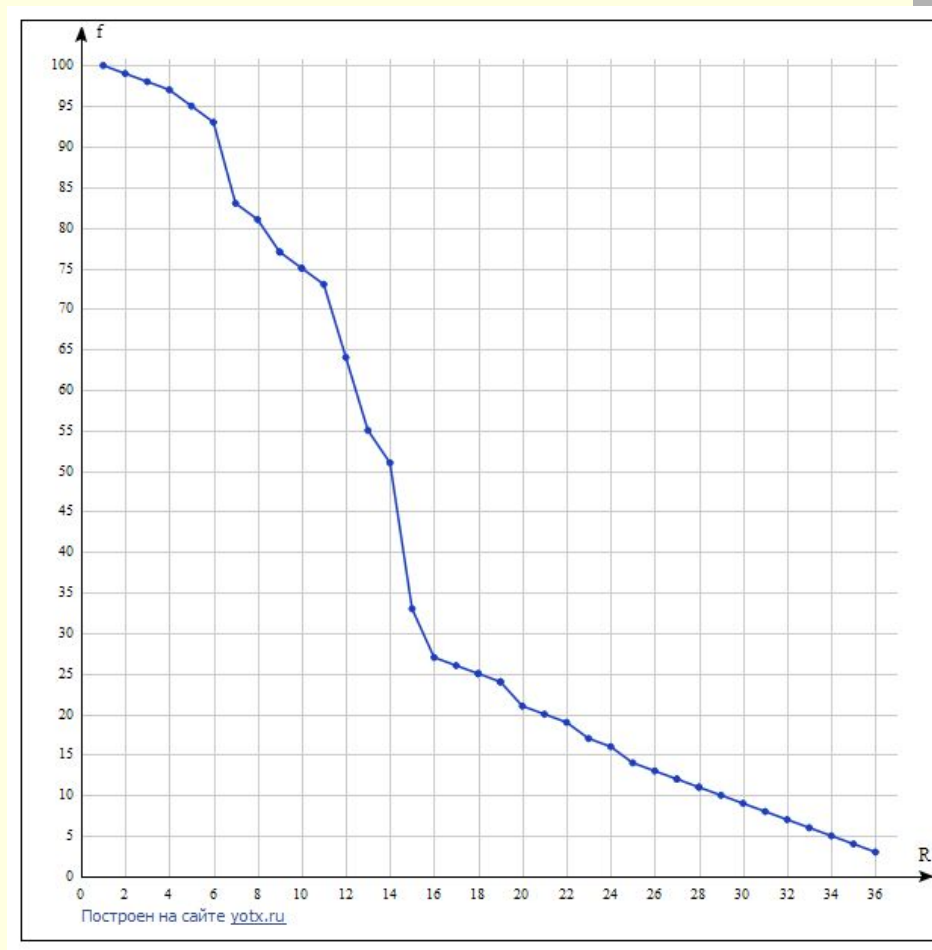
- анализ содержания текста с автоматическим формированием семантической сети с гиперссылками - получения смыслового портрета текста в терминах основных понятий и их смысловых связей;
- анализ содержания текста с автоматическим формированием тематического древа с гиперссылками - выявления семантической структуры текста в виде иерархии тем и подтем;
- смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста;
- автоматическое реферирование текста - формирования его смыслового портрета в терминах наиболее информативных фраз;
- кластеризация информации - анализ распределения материала текстов по тематическим классам;
- автоматическая индексация текста с преобразованием в гипертекст;
- ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;
- автоматическое/автоматизированное формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

TextAnalyst позволяет осуществлять эффективную семантическую обработку текстов с извлечением ключевых слов и выражений.

ТАБЛИЦА КЛЮЧЕВЫХ СЛОВ (TextAnalyst)

100 ERP	11 Качественное
99 CRM	11 применением
99 ERP II	11 цель
99 SCM	10 прозрачности
99 бизнес	10 сотрудничества
99 бизнес-процессов	9 гибкости
99 бизнес-цепочки	8 Planning
99 клиентов	8 архитектуры ERP II
99 компании	8 обработки
99 охватывающими	8 распространяются
99 партнеров	7 системы управления
99 предприятия	7 технологию ERP
99 процессов	7 экономике
99 распределяющее	6 ПО
99 управления	6 сетевое
99 ценности	6 сотрудники
98 поставщика	5 ERP II много технических
97 ERP-системы	5 Systems
97 повышает	5 взаимодействия
95 приложений	5 внедрения системы ERP II
93 ограничиваются	5 возможностей недостаточно для управления
93 участников	5 долгосрочных внутренних и внешних отношений
83 повышающие ценность	5 инерция
81 производственные	5 конкурентных преимуществ
77 изменения	5 обслуживания клиентов
75 поддержки	5 потребностях
73 культуру	5 технологической
64 Management	5 управления запасами
55 Relationship Management	5 Эволюция
51 способно	4 Gartner Group
33 заказов	4 бэк-офисное ERP
27 розничный	4 достижения
26 представляют	4 модулями
25 PRM	4 офисные
25 охватывающей всю бизнес-цепочку	4 получают полную информацию
24 ориентированы на повышение	4 привести
21 ориентировано на клиентов	4 Термин ERP
20 участников бизнес-цепочки	4 фронт-офисными
19 ценных клиентов	3 ERP-систему
17 автономными	3 MRP
16 изменения бизнес-процессов	3 времени
16 создание «распределенного	3 достижения этой цели
14 CRM – SCM	3 инновационных
14 внедрения ERP II	3 методика ERP
14 поколения	3 прозрачность доступа к информации
14 улучшения	3 создавалась
13 интеллектуальным	3 способствует
13 культуре и бизнес	3 точках взаимодействия
13 сложнее	3 фабрики
12 продаж	

График зависимости частоты вхождения слова от ранга



ФОРМУЛИРОВКА ПОИСКОВОГО ЗАПРОСА (TextAnalyst)

- Поисковой запрос с использованием ключевого слова «ИЛИ» (or):

The screenshot shows a Google search interface with the following elements:

- Search Bar:** Contains the query "erp or crm or scm or бизнес or бизнес-процессов or бизнес-цепочки or кли".
- Navigation:** Includes tabs for "Поиск", "Картинки", "Видео", "Новости", "Карты", and "Книги".
- Results:** Shows approximately 759 results. The top results are:
 - Result 1:** Title: "Эволюция технологий ERP-систем". URL: "capri.urfu.ru/IntelligentE/vershiny.htm". Description: "Предметные подобласти SCM и CRM — такие, как APS (Advanced ... В ERP II процессы цепочек поставок компании интегрированы с ... но также и с бизнес-процессами своих партнеров, клиентов, поставщиков и дистрибьюторов. ... промежуточного ПО класса EAI (интеграции приложений предприятия)."
 - Result 2:** Title: "Информационные технологии в бизнесе Часть 1.pdf". URL: "elib.bsu.by/.../Информационные%20технологии%20в%20бизнесе...". Description: "1) управляющие бизнес-процессы, управляющие функционированием сис- ... ся в электронных сетях, охватывающих отдельное предприятие или ... мизирующими работу с клиентами и поставщиками (CRM-системы), а также ... SCM (Supply Chain Management) – системы управления цепочками поста-".
 - Result 3:** Title: "4.4.1 Приложения для предприятий". URL: "www.e-uni.ee/e-kursused/eucip/...vk/441___html". Description: "бизнес-приложения, т.е. специфические приложения для поддержки основной ... в том числе универсальные приложения процессов поддержки, общего ... Системы управления взаимоотношений с клиентами (CRM) ... клиентов, охватывающей все предприятия, участвующие в цепочке поставок Управление ..."
- Filters:** On the left side, there are filters for "Все страны" (Belarus), "На всех языках" (Russian), "За всё время" (Time range: 1 hour, 24 hours, 1 week, 1 month, 1 year), and "Все результаты" (Exact match).

ЗАКЛЮЧЕНИЕ

- В ходе выполнения данной лабораторной работы я ознакомился с одной из методик выбора ключевых слов поискового запроса, применил ее для поиска необходимого документа, т.е. осуществил профессиональный поиск информации. Также познакомился с работой программы-экстрактора TextAnalyst, выполнил с ее помощью анализ текста и, на основе полученных данных, построил поисковой запрос.