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 $C = f*R \sim const$.

График зависимости частоты слова f от его ранга R.

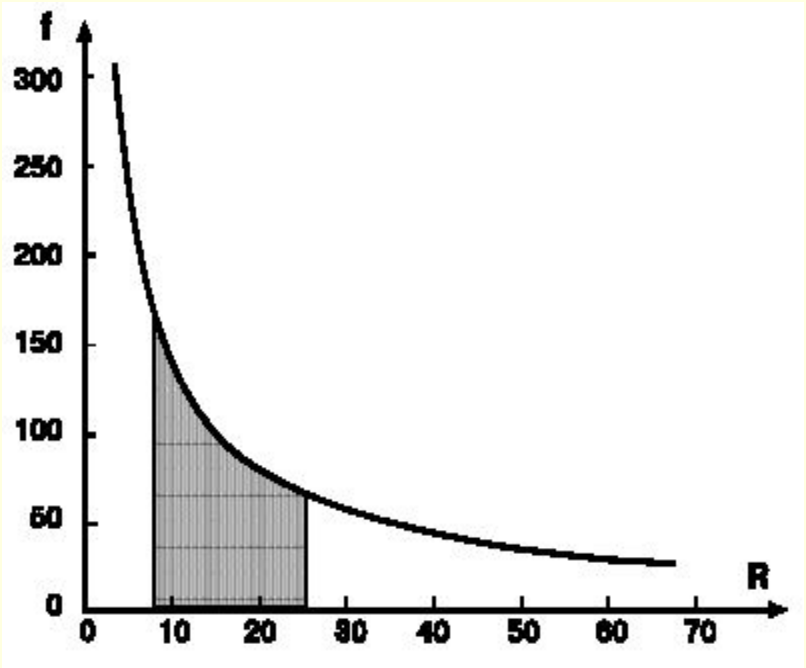


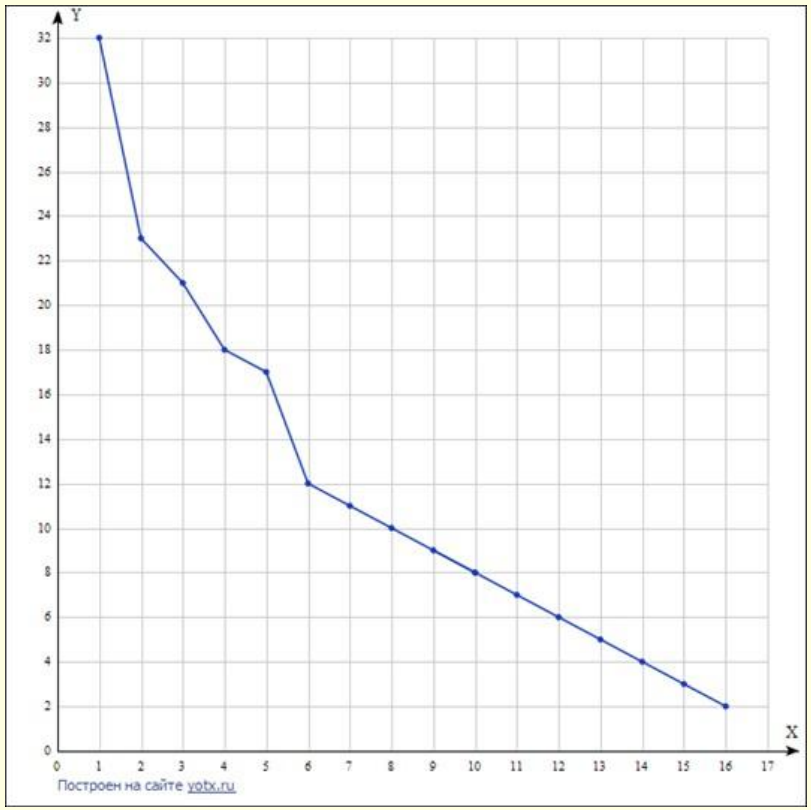
ТАБЛИЦА КЛЮЧЕВЫХ СЛОВ (ручной поиск)

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

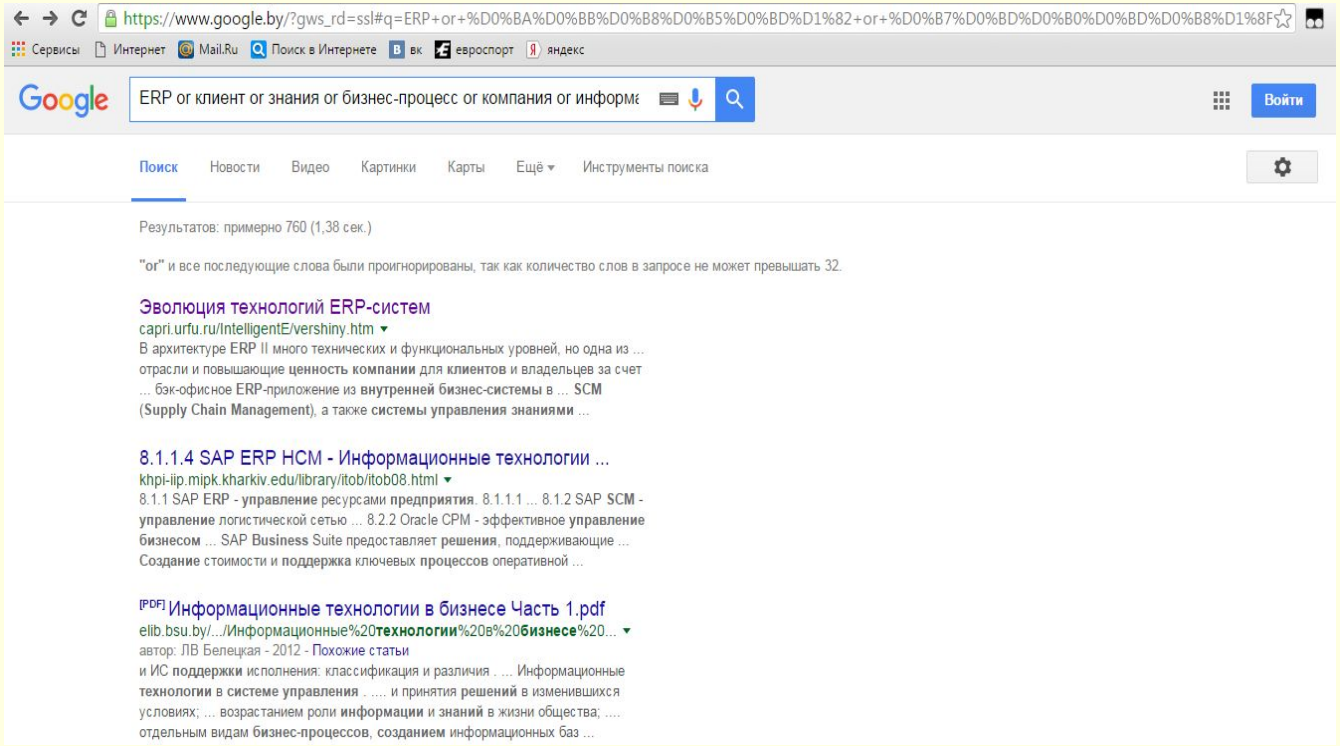
График зависимости частоты вхождения слова от ранга

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.



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.

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

ТАБЛИЦА КЛЮЧЕВЫХ СЛОВ (TextAnalyst)

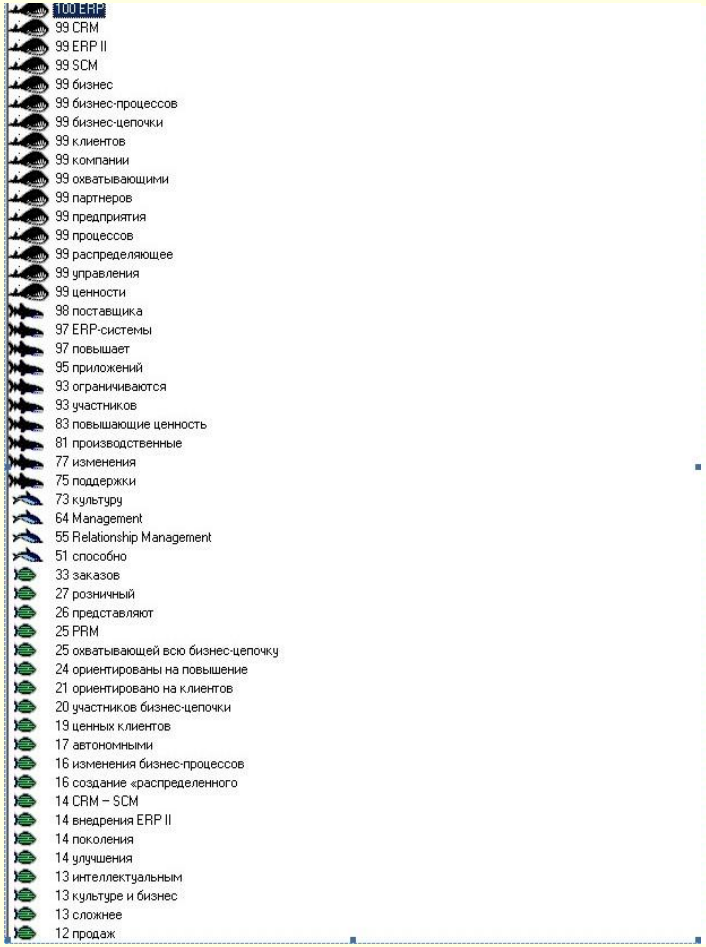
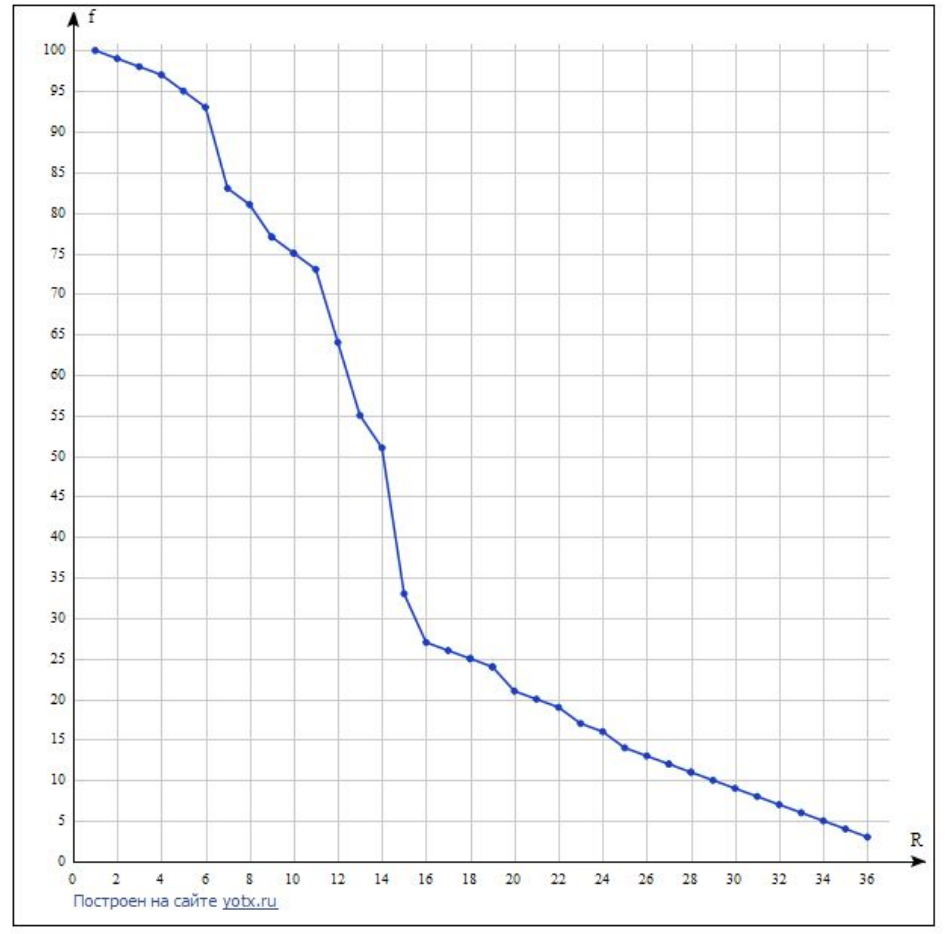
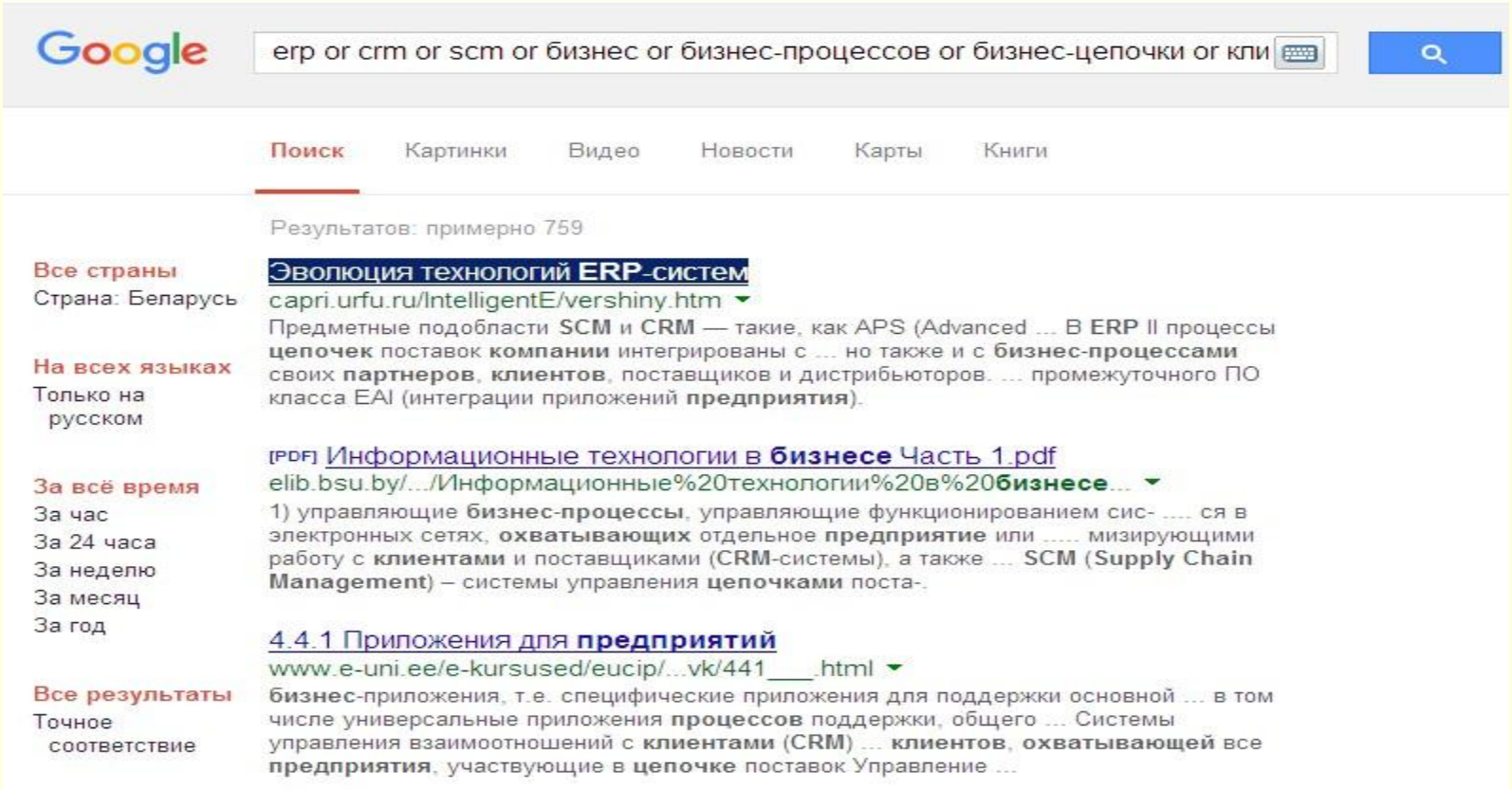


График зависимости частоты вхождения слова от ранга



ФОРМУЛИРОВКА ПОИСКОВОГО ЗАПРОСА (TextAnalyst)

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



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.