Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра информационных технологий автоматизированных систем

Лабораторная работа №2

«Технология анализа текста и извлечения ключевых слов»

Вариант 4

|  |  |
| --- | --- |
| Проверил:  Стригалев Л.С. | Выполнил:  студент группы 120602  Будный Р. И. |

Минск 2013

# 1. Цель работы

Практическое освоение технологии анализа текста, извлечения ключевых слов и профессионального поиска информации.

# 2. Краткая характеристика использованных методов построения поискового запроса

Для грамотного построения запросов поисковой машине, а также для правильного выбора ключевых слов поиска используются законы Зипфа.

Первый закон Зипфа описывается при рассмотрении некоторого произвольного текста. Выписывают все различающиеся слова данного текста в виде с, с,…,с,…,с, где с- i-ое слово, не совпадающее ни с каким другим словом в данном множестве. Для каждого из этих слов подсчитывается количество его повторов в тексте. В результате получают f, f,…,f,…,f, где f- количество повторений i-го слова в тексте, названное Зипфом *частотой слова*. Далее располагают слова в порядке убывания их частот и ставят в соответствие числа натурального ряда, называющиеся рангами слов (R): слову с максимальной частотой присваивается ранг 1, следующему по частоте — ранг 2 и т.д. При этом если несколько разных слов имеют одинаковые частоты, то они объединяются в один блок. Зипф ввел понятие *вероятности встречи слова* как отношение частоты слова к общему количеству слов в тексте.

На рис.1 приведена зависимость частоты слов f от его ранга R. Кривая представляет собой гиперболу f = C / R.

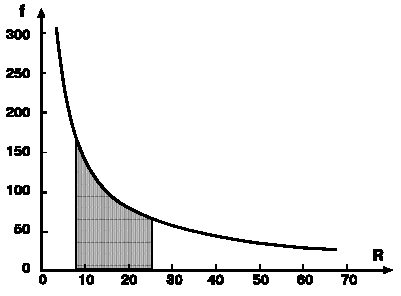


Рис. 1. График зависимости частоты слова f от его ранга R.

Все значимые (ключевые) для текста слова размещаются в средней части графика (выделенная область рис. 1). При этом часто встречающиеся слова (ранг от 1 до 5) являются *вспомогательными* (это предлоги, частицы, местоимения и т.д.). Такие слова называются также «стоп-словами»; поисковая машина при простом поиске их по умолчанию игнорируют, так как эти слова являются «шумом», помехой, которая затрудняет поиск. Редко встречающиеся слова (правая часть графика) также не имеют существенного значения при анализе текста.

Далее выбирают диапазон значений ранга для извлечения ключевых слов. Он не должен быть слишком широким, потому что значимые слова затеряются среди второстепенных; и не должен быть очень узкий, потому что возможна потеря существенных ключевых слов.

Законы Зипфа используются при создании на поисковых серверах базы данных, в которой хранится индексированная информация; при этом учитывается целый ряд факторов, таких как вес слова, его местоположение в документе, морфологические особенности и др. Уточненные законы Зипфа используются также в алгоритмах автоматического распознавания текста программ-экстракторов, которые осуществляют семантический анализ текстов и извлекают ключевые слова и выражения.

**Программы-экстракторы**

Возможность извлечения ключевых слов из текстовых материалов имеется и в текстовом редакторе MS Word (меню «**Сервис**» > «**Автореферат**»**),** однако использование этой возможности дает неудовлетворительные результаты. Из программ-экстракторов, которых в настоящее время не так много, одними из лучших являются **TextAnalyst и RCO Fact Extractor.**

**RCO Fact Extractor —** персональная аналитическая система, предназначенная для автоматического анализа текста и выделения фактов, связанных с интересующими объектами. RCO Fact Extractor позволяет найти в тексте описания фактов заданного типа и извлечь связанную с ними информацию. Основная сфера применения этого программного средства — аналитические задачи из области компьютерной разведки, требующие высокоточного отбора информации по заданным смысловым критериям, например, автоматизированное составление досье на целевые персоны или организации.

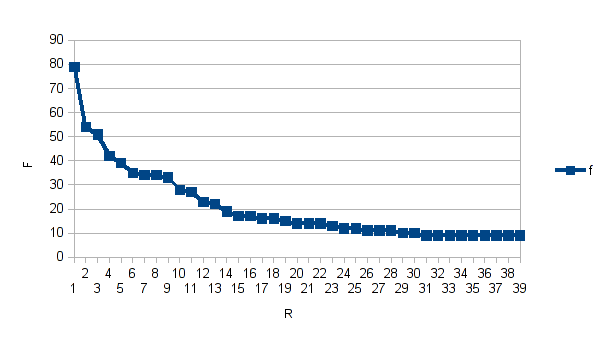
Персональная система автоматического анализа текста **TextAnalyst** предназначена для анализа содержания текстов, смыслового поиска информации и формирования электронных архивов. Он позволяет осуществлять эффективную семантическую обработку текстов с извлечением ключевых слов и выражений. Эта программа может быть использована и для выявления нарушения авторских прав.

# 3. Частоты вхождения и ранги слов текста-источника. График зависимости частоты слова f от его ранга R

В ходе проведения лабораторной работы мной был проанализирован текст под названием «Говорим — ИТ-проект, подразумеваем — расширенный». Были посчитаны частоты вхождения слов и определены ранги слов (Таблица 1). График зависимости частоты слова f от его ранга R представлен на рисунке 2.

|  |  |  |
| --- | --- | --- |
| **Слово** | **f** | **R** |
| В | 79 | 1 |
| Проект | 76 | 2 |
| И | 51 | 3 |
| Что | 42 | 4 |
| Не | 39 | 5 |
| Рп | 35 | 6 |
| На | 34 | 7 |
| С | 34 | 7 |
| Это | 33 | 8 |
| Будет | 28 | 9 |
| По | 27 | 10 |
| Ит | 23 | 11 |
| Можно | 19 | 13 |
| Системы | 17 | 14 |
| Как | 17 | 14 |
| Или | 16 | 15 |
| К | 16 | 15 |
| То | 15 | 16 |
| Фаз | 14 | 17 |
| Компании | 14 | 17 |
| Для | 14 | 17 |
| Его | 13 | 18 |
| О | 12 | 19 |
| Следует | 12 | 19 |
| Случае | 11 | 20 |
| Если | 11 | 20 |
| А | 11 | 20 |
| Её | 10 | 21 |
| Будут | 10 | 21 |
| Здесь | 9 | 22 |
| Быть | 9 | 22 |
| Фазы | 9 | 22 |
| Из | 9 | 22 |
| Может | 9 | 22 |
| От | 9 | 22 |
| Только | 9 | 22 |
| Все | 9 | 22 |
| Всего | 9 | 22 |

Таблица 1 – Частоты вхождения и ранги слов

Рисунок 2 - График зависимости частоты слова f от его ранга R

# 4. Описание результатов поиска документов

На основе полученных данных, я выбрал диапазон ранга слов и сформировал список ключевых слов. Далее, составил поисковой запрос, используя логический оператор «ИЛИ»**,** выбирая диапазон слов и соблюдая порядок следования ключевых слов сформированного списка:

*проект or рп or системы or компании or фазы or ситуации or команда or проблемы*

Результат поиска представлен на рисунке 3. Мы видим, что на сформированный запрос поисковая машина выдала нам на первых местах наиболее релеватные документы, правда, искомого документа в выдаче не оказалось.

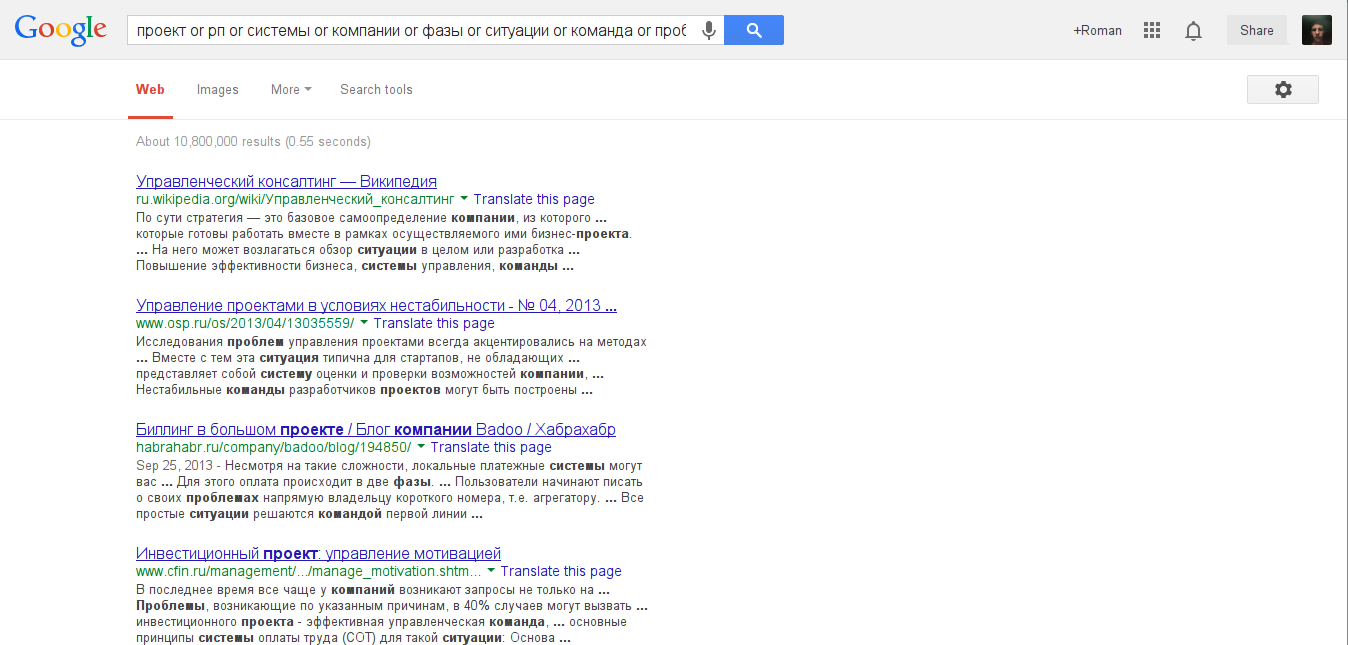


Рисунок 3 – Поисковой запрос без использования программных средств с оператором ИЛИ

Также сформируем запрос без использования логического оператора «ИЛИ»**:**

*проект рп системы компании фазы ситуации команда проблемы*

Результат представлен на рисунке 4. По содержанию выдача совпадает с выдачей предыдущего запроса.

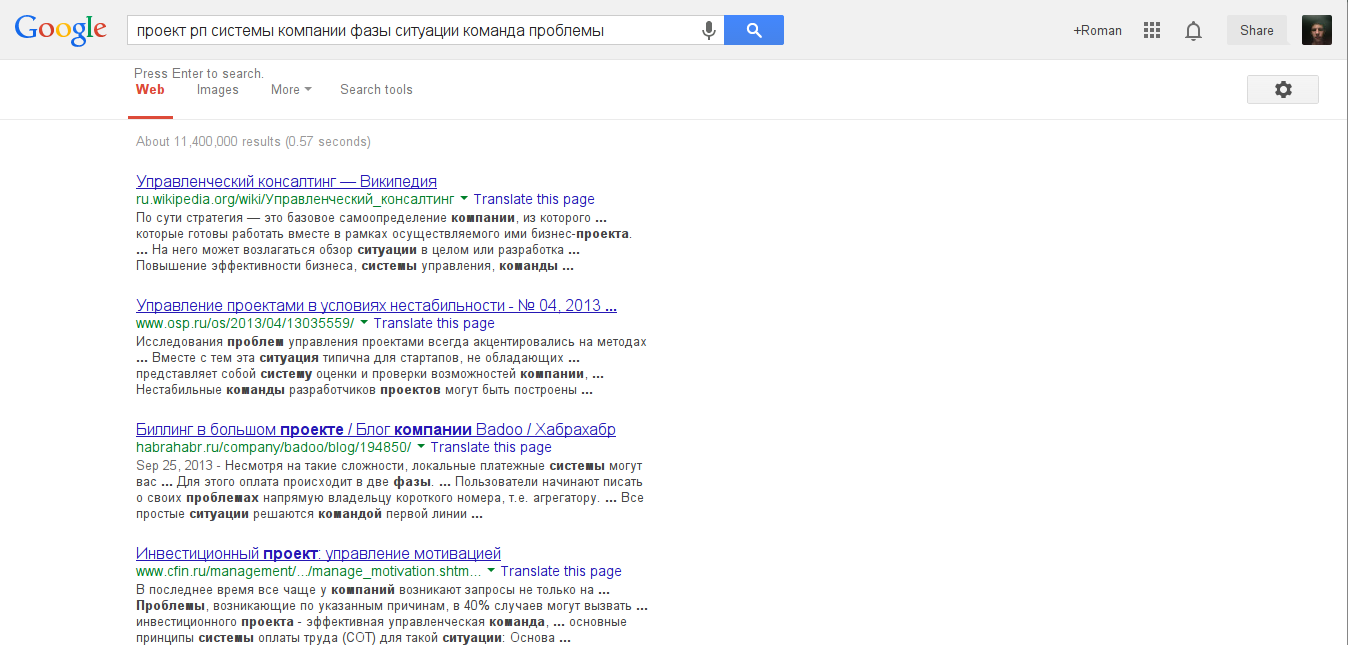


Рисунок 4 – Поисковой запрос без использования программных средств и без оператора ИЛИ

Далее составим запрос на основе малого диапазона ключевых слов:

*проект рп системы компании фазы*

Результат виден на рисунке 5. Выдача поисковой системы не отвечает нашим стремлениям, т.к. на первых страницах находится информация не по теме. Следовательно, для успешного поиска нужно использовать широкий диапазон ключевых слов.

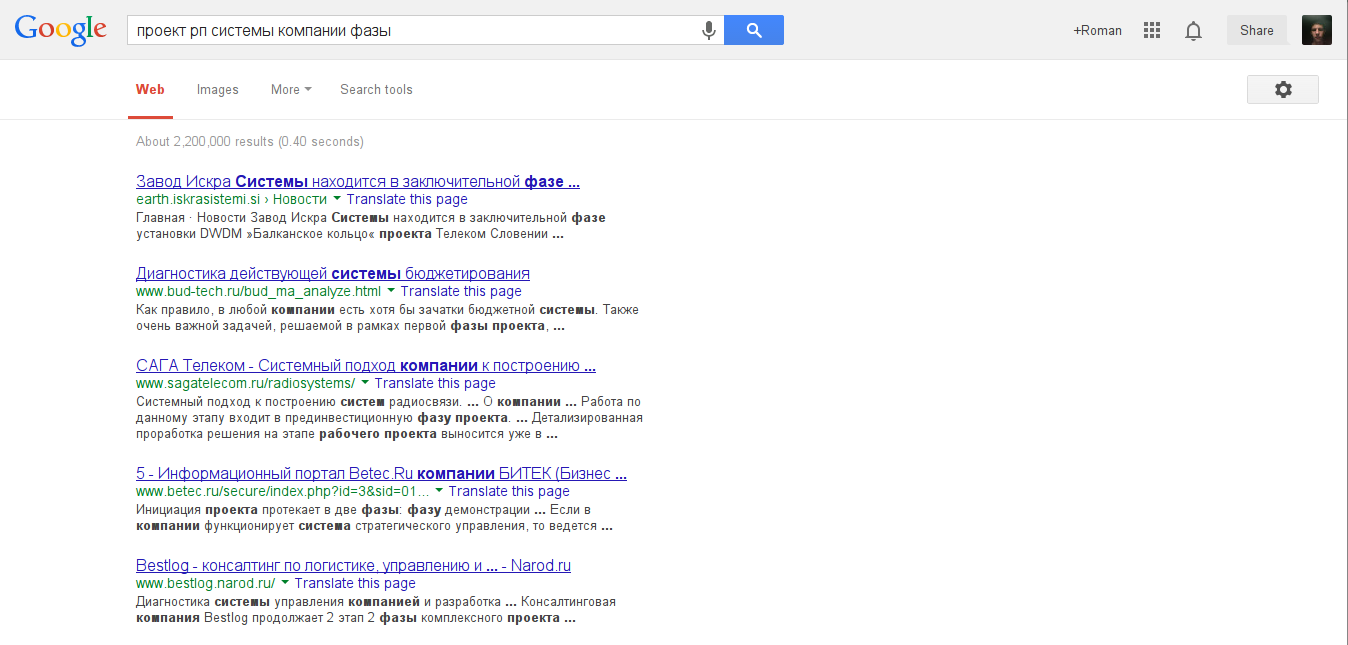


Рисунок 5 - Поисковой запрос без использования программных средств и с малым диапазном слов

Теперь построим поисковый запрос с использованием системы TextAnalyst по томуже тексту-источнику.В результате анализа текста (рисунок 6), сформируем список наиболее значимых понятий (таблица 2).График зависимости частоты слова f от его ранга R представлен на рисунке 7.

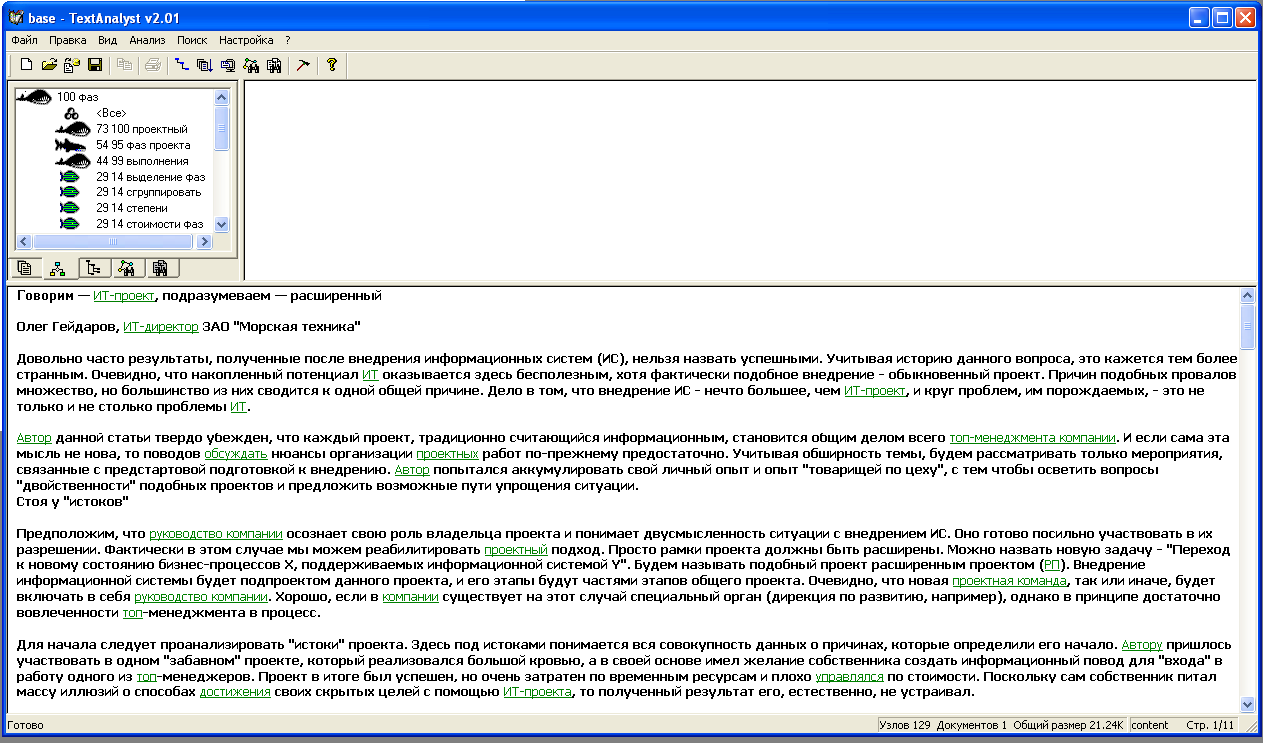


Рисунок 6 – Окно с результатом анализа текста в системе TextAnalyst

|  |  |  |
| --- | --- | --- |
| Слово | f | R |
| РП | 35 | 1 |
| фаз | 26 | 2 |
| проектный | 24 | 3 |
| ИТ | 23 | 4 |
| компании | 15 | 5 |
| процесс | 11 | 6 |
| бизнес | 9 | 7 |
| ИТ-проект | 9 | 7 |
| представляется | 9 | 7 |
| автор | 8 | 8 |
| получить | 8 | 8 |
| управления | 7 | 9 |
| проектной команды | 6 | 10 |

Таблица 2 – Список основных понятий

Рисунок 7 - График зависимости частоты слова f от его ранга R

Далее составим поисковой запрос, используя логический оператор «ИЛИ» и соблюдая порядок следования ключевых слов сформированного списка:

*РП or фаз or проектный or ИТ or компании or процесс or ИТ-проект or представляется or автор or получить*

Результат представлен на рисунке 8. Необходимый документ содержится на первой странице.

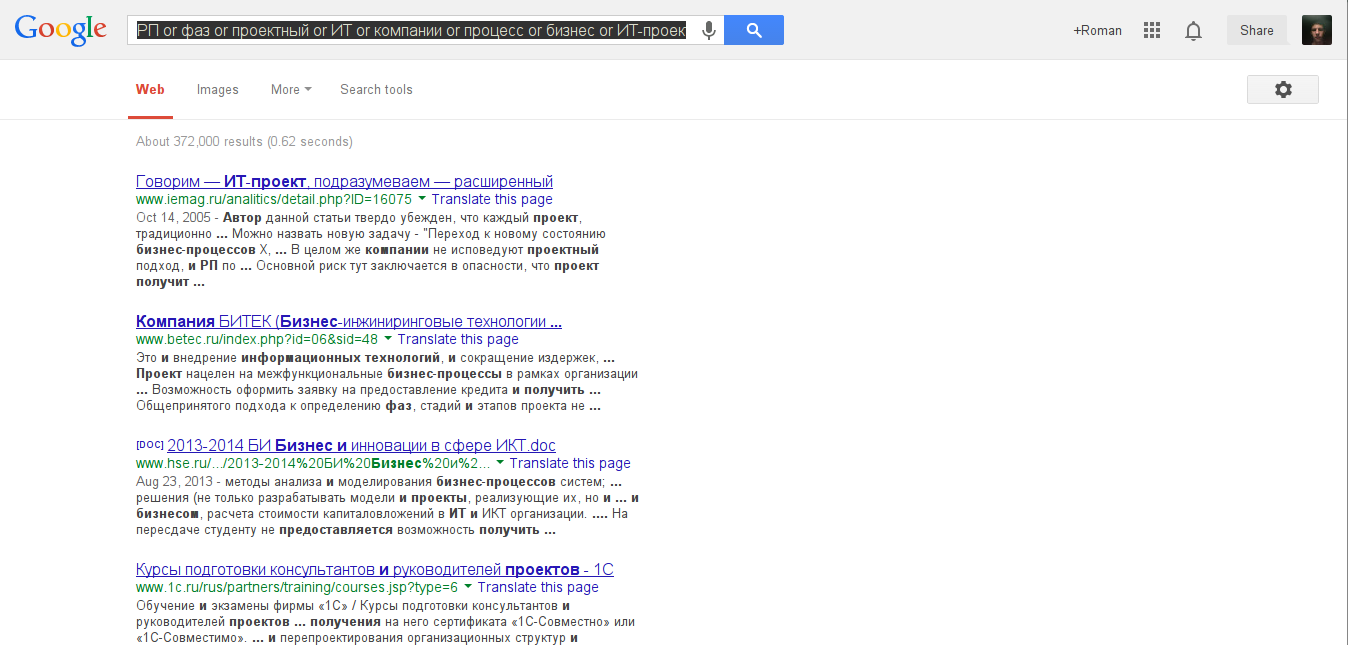


Рисунок 8 – Поисковой запрос с использованием аналитической системы TextAnalyst с оператором ИЛИ

Составим поисковой запрос без использования оператора ИЛИ:

*Предприятия членов группы Координационный комитет подразделений процесса выполнение представлять рамках внедрения повышению*

Результат представлен на рисунке 9. Результат аналогичен результату предыдущего запроса.

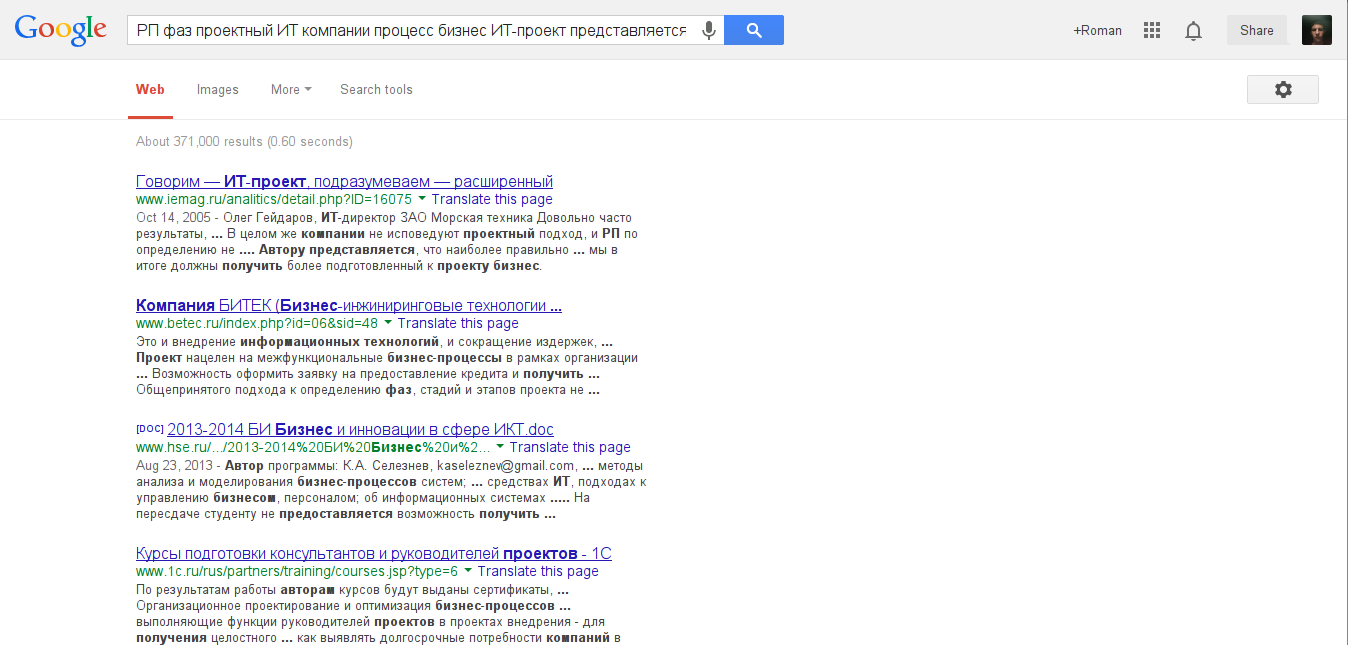


Рисунок 9 - Поисковой запрос с использованием аналитической системы TextAnalyst без оператора ИЛИ

Сформируем запрос с небольшим диапазоном слов:

*РП фаз проектный ИТ компании*

Результат можем наблюдать на рисунке 10. Искомый документ на первой странице поисковой выдачи не появился, что говорит о недостаточной точности поисковой составленной поисковой фразы.

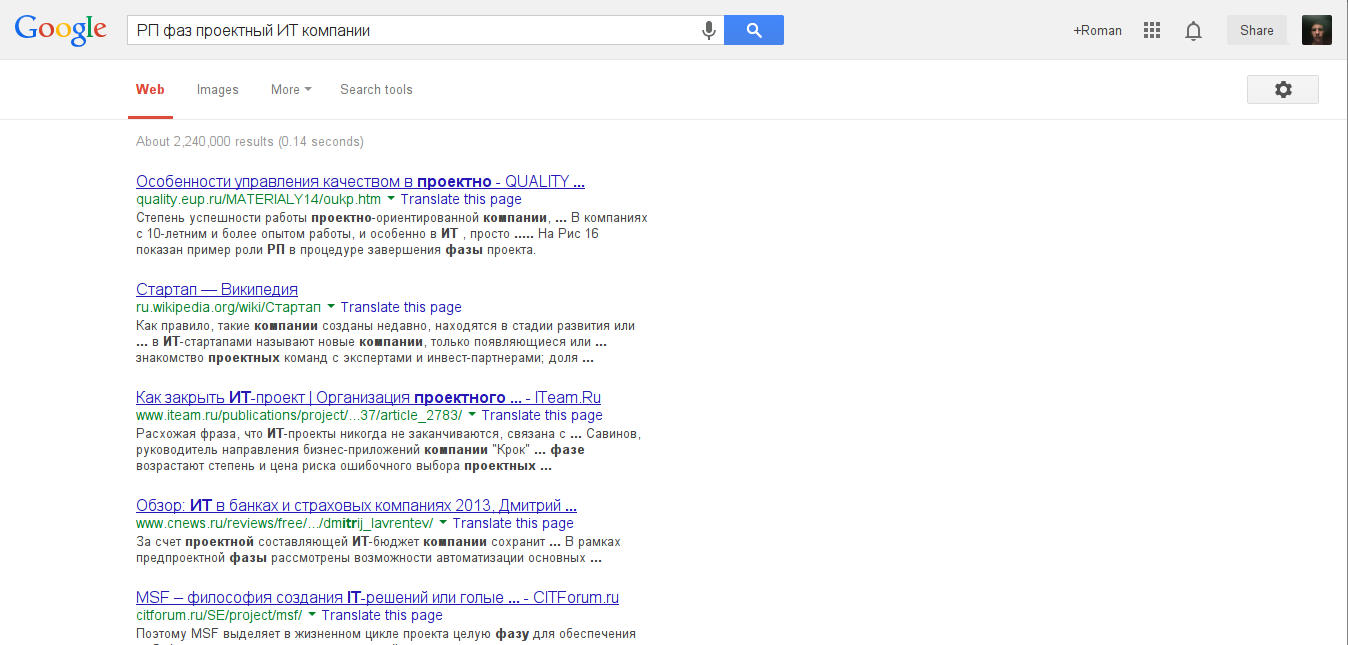


Рисунок 10 - Поисковой запрос с использованием аналитической системы TextAnalyst и с малым диапазоном слов

# 5. Анализ полученных результатов

Графики зависимости частоты слова f от его ранга R для двух способов анализа текста схожи, но использование программы TextAnalyst дает более точные результаты.

Поиск с помощью аналитической системы осуществляется эффективнее. TextAnalyst формирует сеть основных понятий, несущих основную смысловую нагрузку. Эта сеть достаточно полно описывает смысл анализируемого текста, а также позволяет отбросить несущественную информацию и представить содержание в сжатом виде.

Рисунок 11 – Графики зависимости частоты слова f от его ранга R для двух способов формирования запросов

# 6. Вывод

В ходе лабораторной работы я освоил освоил применение первого закона Зипфа для анализа текста, извлечения ключевых слов и последующего поиска информации в сети Интернет. Сравнил эффективность анализа текста вручную с использованием вышеупомянутого закона и применение специализированного ПО для анализа текста.