**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ**

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И УПРАВЛЕНИЯ

**Лабораторная работа № 2**

**технология анализа текста и извлечения ключевых слов**

|  |
| --- |
| Выполнил: |
| Студент гр 322441  Калинин В.А. |
| Проверил: |
|  |

Минск 2023

Содержание

1)Цель работы.

2)  Краткая характеристика использованных методов построения поискового запроса.

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

4) Описание результатов поиска документов.

5)  Анализ полученных результатов.

6) Выводы по работе.

**1) Цель работы**

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

**2) Краткая характеристика использованных методов построения поискового запроса.**

Джордж Зипф установил, что все тексты подчиняются общим закономерностям, и сформулировал в 1946—49 гг. несколько законов, которые нашли применение в технологии поиска информации.

Для ознакомления с положениями первого закона Зипфа введем, следуя Зипфу, необходимые терминологические определения. Рассмотрим некоторый произвольный текст. Выпишем все различающиеся слова данного текста в виде сhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-ZDwjav.png, сhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-X2jrpM.png,…,сhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-bore4K.png,…,сhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-zTKCHb.png, где сhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-q0YgQF.png- i-ое слово, не совпадающее ни с каким другим словом в данном множестве. Для каждого из этих слов подсчитаем количество его повторов в тексте. В результате получим fhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-iUfsKf.png, fhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-krswuv.png,…,fhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-Jkt3oY.png,…,fhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-0Y8kUX.png, где fhttps://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-PDzevT.png- количество повторений i-го слова в тексте, названное Зипфом **частотой слова** (в данном случае i-го слова). Далее Зипф, располагая слова в порядке убывания их частот, поставил им в соответствие числа натурального ряда, назвав эти числа **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. Наконец Зипф ввел понятие **вероятности встречи слова** как отношение частоты слова к общему количеству слов в тексте (в математической статистике такое отношение называется частотой события; а в справочной системе поисковых машин — относительной частотой).

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 https://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-uii__9.pngR https://studfile.net/html/1549/349/html_mnRhRLIA9H.ViMp/img-YdBylk.png const.

Комментарий по поводу круглых скобок. Если левую и правую часть этого равенства разделить на общее количество слов в тексте, то равенство не нарушится, но его правая часть будет не что иное, как произведение вероятности встречи слова на его ранг.

Законы Зипфа используются при создании на поисковых серверах базы данных, в которой хранится индексированная информация. Они же используются и для оценки релевантности (степени соответствия) документов в процессе поиска. Уточненные законы Зипфа используются также в алгоритмах автоматического распознавания текста программ-экстракторов, которые осуществляют семантический анализ текстов и извлекают ключевые слова и выражения.

При выборе ключевых слов поискового запроса **без использования программных** средств необходимо выполнить следующее:

1. Удалить (или вычеркнуть) из выбранного текста все стоп-слова и вычислить частоту вхождения каждого из оставшихся слов. Заметим, что так работает поисковая машина. В учебных целях стоп-слова можно и не вычеркивать.

2. Выписать слова в порядке убывания их частоты вхождения f и присвоить словам ранги (см. п. 2.1). В учебных целях рекомендуется построить график зависимости частоты слова f от его ранга R.

3. Выбрать, руководствуясь здравым смыслом, диапазон значений ранга слов. При этом следует помнить, что при слишком широком диапазоне значимые слова затеряются среди второстепенных слов, а при очень узком значимые слова могут просто потеряться.

4. Сформировать, используя выбранный диапазон ранга слов, список ключевых слов. Достаточно взять 10-20 слов.

5. Составить поисковый запрос, используя логический оператор «ИЛИ»и соблюдая порядок следования ключевых слов сформированного списка. Напомним (см. Лаб. раб. № 1), по умолчанию поисковая машина использует логический оператор «И»**,**так что если оператор «ИЛИ**»**в поисковом запросе не указан, то при значительном количестве ключевых слов в запросе результаты поиска могут оказаться нулевыми.

На сформированный таким образом запрос поисковая машина может выдать несколько миллионов страниц, но, поскольку поисковая машина, как правило, ранжирует результаты поиска, то на первых страницах окажутся наиболее релевантные документы.

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

По методу, приведённому в предыдущем пункте для выбора ключевых слов поискового запроса, был рассмотрен текст полученного варианта (6 вариант).

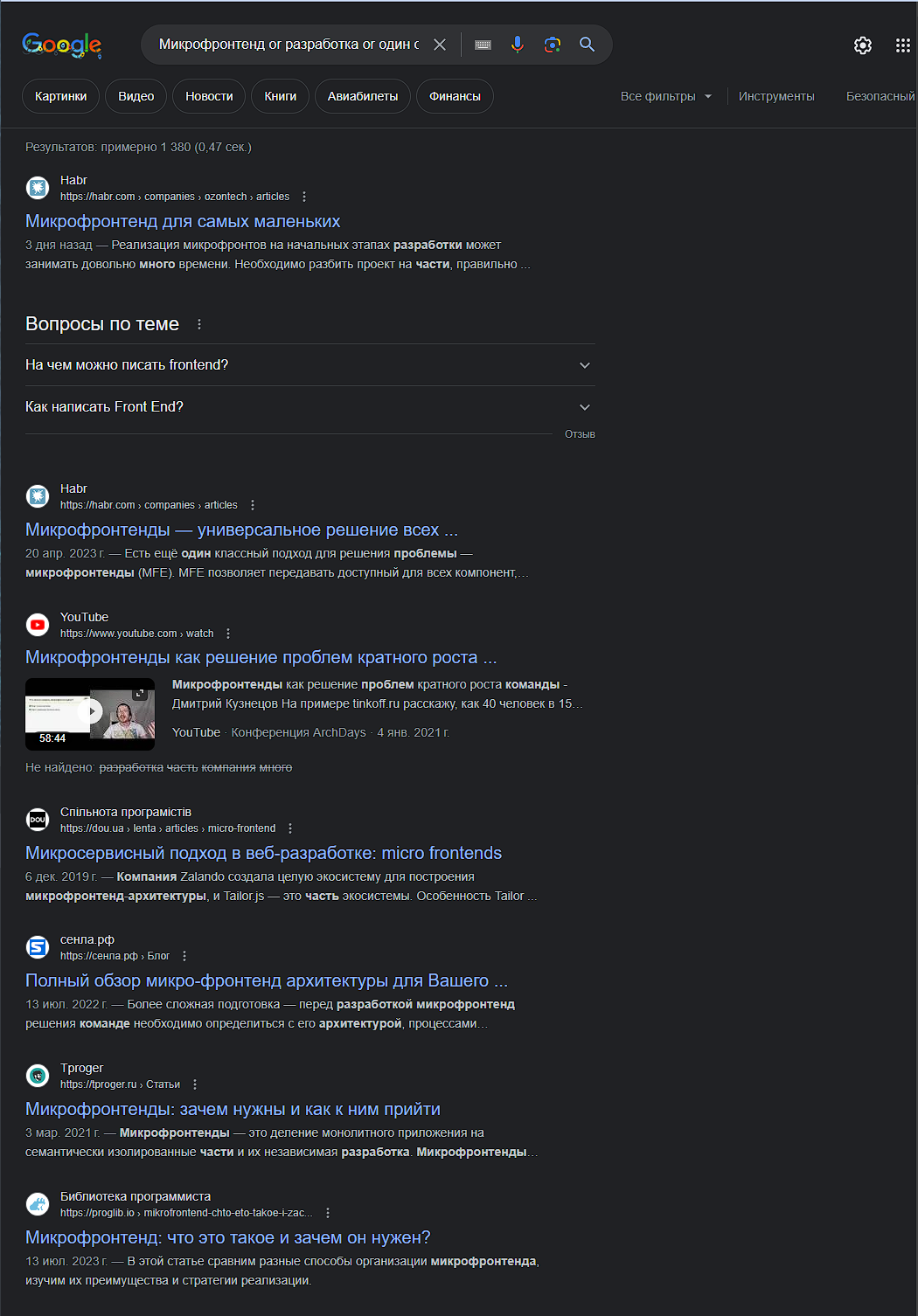
Сначала был проведен ручной поиск ключевых слов в приведенном тексте. Потом была проведена аналогичная работа с помощью программы TextAnalyst ( https://advego.com/text/seo/ ).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| № | СЛОВО | ЧАСТОТА | РАНГ | № | СЛОВО | ЧАСТОТА | РАНГ |
| 1 | приложение | 25 | 1 | 20 | оно | 5 | 10 |
| 2 | микрофронтенд | 19 | 2 | 21 | панель | 5 | 10 |
| 3 | разработка | 18 | 3 | 22 | разный | 5 | 10 |
| 4 | один | 10 | 4 | 23 | разрабатывать | 5 | 10 |
| 5 | часть | 10 | 4 | 24 | selectel | 4 | 11 |
| 6 | команда | 9 | 5 | 25 | блок | 4 | 11 |
| 7 | все | 8 | 6 | 26 | использовать | 4 | 11 |
| 8 | проблема | 7 | 7 | 27 | может | 4 | 11 |
| 9 | архитектура | 6 | 8 | 28 | писать | 4 | 11 |
| 10 | компания | 6 | 8 | 29 | продукт | 4 | 11 |
| 11 | много | 6 | 8 | 30 | процесс | 4 | 11 |
| 12 | монорепозиторий | 6 | 9 | 31 | работать | 4 | 11 |
| 13 | несколько | 6 | 9 | 32 | смысл | 4 | 11 |
| 14 | библиотека | 6 | 9 | 33 | фронтенд | 3 | 12 |
| 15 | бэкенд | 5 | 10 | 34 | всегда | 3 | 12 |
| 16 | деление | 5 | 10 | 35 | инструмент | 3 | 12 |
| 17 | микросервисный | 5 | 10 | 36 | код | 3 | 12 |
| 18 | монолит | 5 | 10 | 37 | микросервис | 3 | 12 |
| 19 | нужно | 5 | 10 |  |  |  |  |

**4. Описание результатов поиска документов.**

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

Микрофронтенд or разработка or один or часть or команда or все or проблема or архитектура or компания or много or монорепозиторий or несколько or библиотека



Мы нашли 1380 сайтов удовлетворяющих нашему запросу. Страницы задания (вариант 6) не оказалось не только на первом месте, но и в первой сотне.

**5.Анализ полученных результатов.**

Был проведен поиск информации на сформированные нами запросы при использовании слов с наибольшей частотой и большими рангами в начале при помощи ручной технологии, в результате поисковая машина выдала достаточное огромное, но при этом достаточно одинаковое количество страниц. Также можно отметить, что поисковая машина, как правило, ранжирует результаты поиска, и на первых страницах у нас оказались наиболее релевантные документы.

Мы по-разному задавали поиск, состоящий из слов с наибольшей частотой. В итоге у нас получались достаточно близкие результаты. И в итоге наша статья (текст – источник) практически никогда не была в десятке первых найденных ссылок.

Работа произведена и её можно считать неуспешной. Мы получили не тот результат, который хотели получить. Виной всему современные методы генерации статей с бОльшим количеством ключевых слов, которые будут попадаться в выдаче гораздо выше оригинальной статьи.

**6. Выводы по работе**

1.Ознакомились с основными принципами выбора ключевых слов (на основе закона Зипфа).

2.Ознакомились с программами-экстракторами (**TextAnalyst и RCO Fact Extractor**).

3.При помощи **TextAnalyst** осуществили эффективную обработку текста с извлечением ключевых слов и выражений.

4. Просчитали ранги слов и построили график зависимости частоты от ранга.

5. Составили поисковой запрос при помощи логического оператора «ИЛИ».

6. Пришли к выводу, что современные методы генерации статей позволяют поднимать их выдаче и использовать закон Зипфа весьма умело.