# Разработка и реализация алгоритма извлечения из текста географических названий, отражающих содержание документа

Выполнила: Ыдырыс Жулдызай,
гр. 7203

Науч.руков.: д.т.н., доцент, с.н.с. ИВТ СО РАН
В.Б.Барахнин

**Были поставлены следующие задачи:**

1.Создание тезауруса географических названий

2. Пополнения тезауруса словоформами географических названий

3. Выявления омонимов и многозначных названий

4. Извлечения из текста документа названий, входящих в соответствующий тезаурус

Тезаурус (от греческого - сокровище) – многозначный термин. Наиболее адекватный русский перевод слова «тезаурус» – это «мир знаний и интересов». Например, «мир знаний и интересов художника – тезаурус художника», «мир знаний и интересов бизнесмена – тезаурус бизнесмена» или иначе говоря, тезаурус предметной области «Бизнес». и т.п. Тезаурус - полный систематизированный набор данных по какой-либо предметной области, позволяющий человеку или компьютеру в ней ориентироваться.

## Структура тезауруса

• нормализованное наименование географического объекта;

• географические координаты;

• административный статус населенного пункта;

• соответствующий индекс (территориальное типовое деление);

•административно-территориальная привязка географического объекта (наименование субъекта РФ, наименование административного района);

•принятые на текущий момент наименования географических объектов (иные варианты названий, сокращенные названия);

• источники данных и другие примечания.


Рисунок 1 – Структура тезауруса.

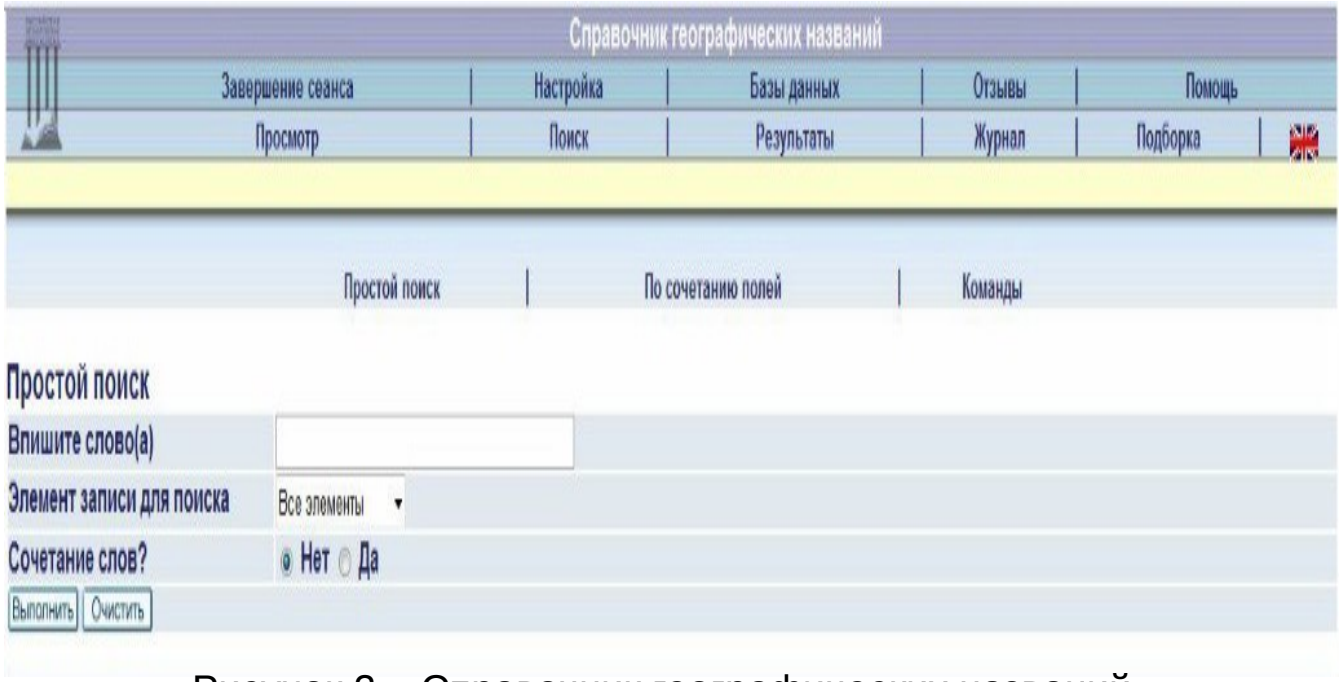



Рисунок 2 – Справочник географических названий РГБ

В справочнике географических названий РГБ подготовлены данные более чем о 140 тыс. географических объектов России (64 региона РФ) и загружены нормативные записи примерно для 80000 географических объектов.


Рисунок 3-Географические названия полученные из словаря Российской государственной библиотеки

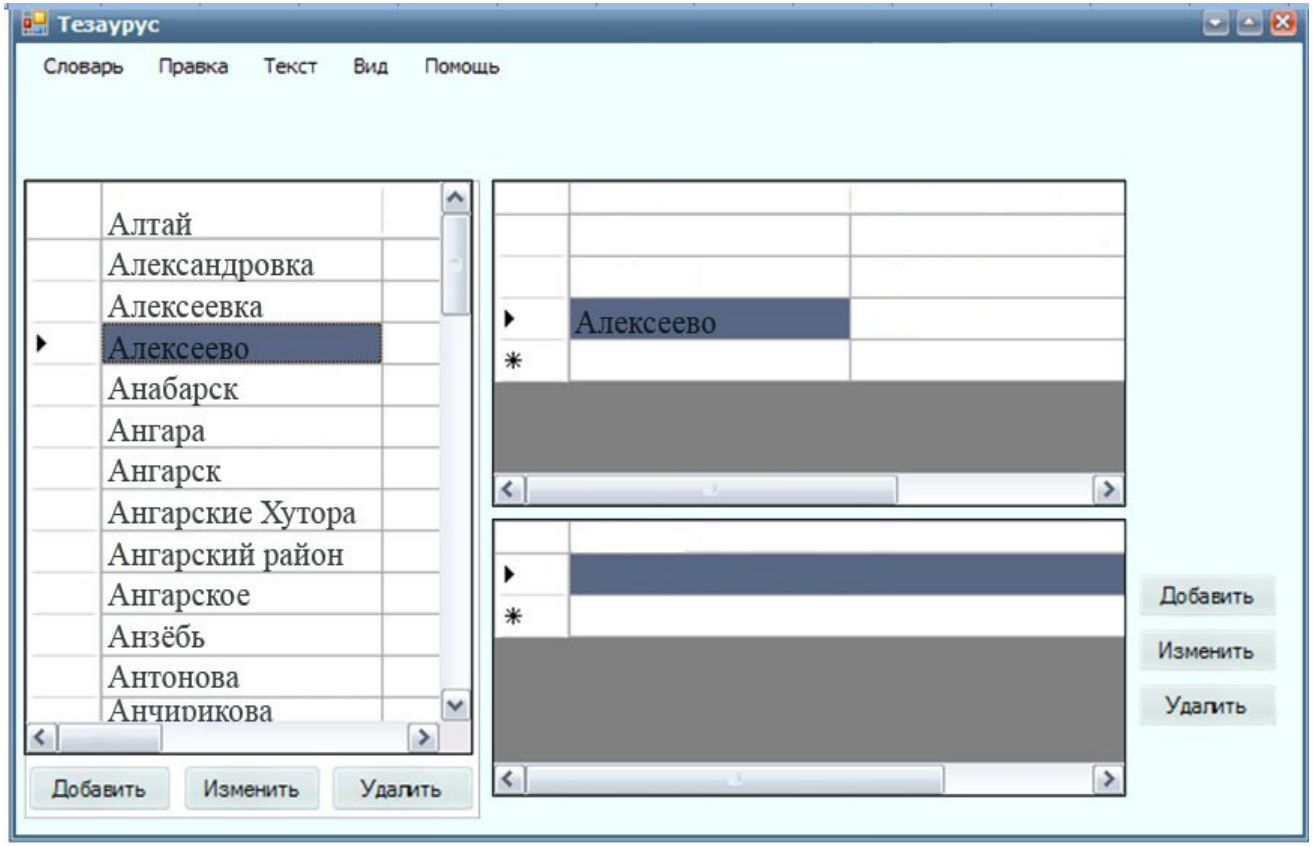



Рисунок 4-Графический интерфейс редактора тезауруса.

Тезаурус можно дополнить синонимами официальных названий: например, название города Санкт-Петербург в разных источниках может быть употреблен как Петербург , Питер, северная столица или же город на Неве.

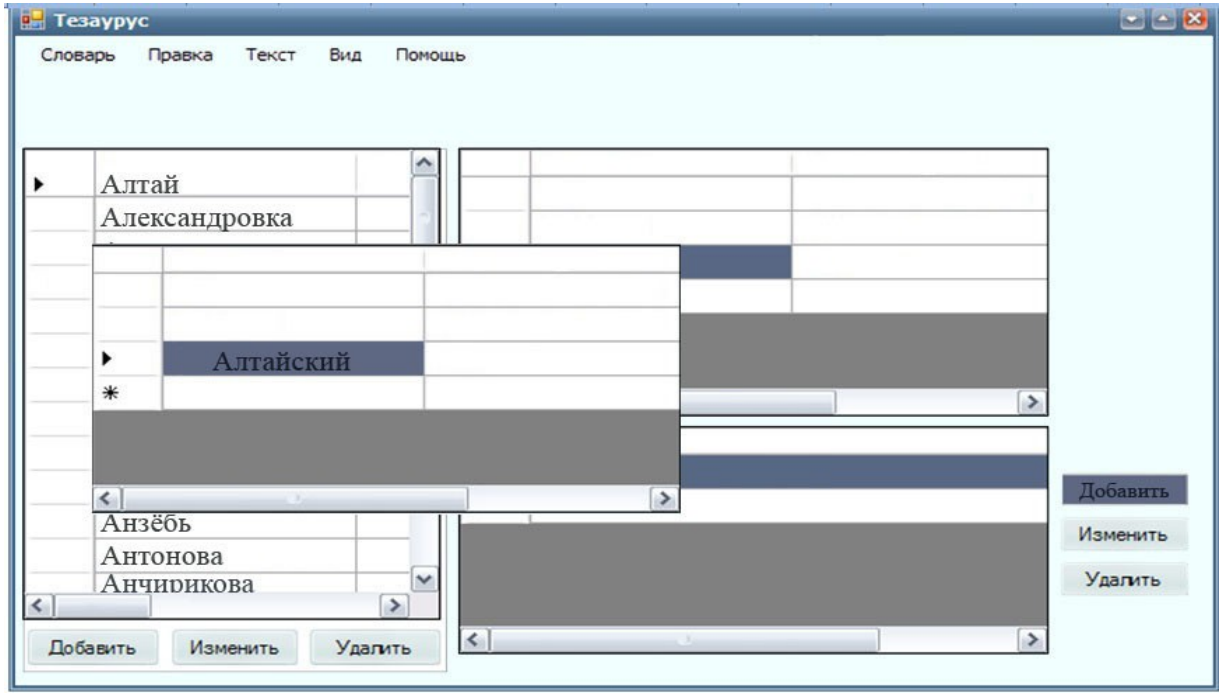



Рисунок 5-Добавление новых названий в тезаурус.

## Пополнение лексического словаря словоформами географических названий

В русском языке слова изменяются по родам и числам. По этому нужно учитывать морфологию.

| Род, число: | женский род | |
|---|---|---|
| кто, что? | Москва | Москвы |
| кого, чего? | Москвы | Москов |
| кому, чему? | Москве | Москвам |
| кого, что? | Москву | Москвы |
| кем, чем? | Москвой | Москвами |
| о ком, о чём? | о Москве | о Москвах |
| где? | в Москве | в Москвах |

Рисунок 6-Склонение названия «Москва»

Флективный класс – класс слов, имеющих одинаковые признаки склонения. То есть слова, которые в определенной форме имеют одинаковые окончания. Характеризуется набором признаков или словом-представителем.

В случае добавления в словарь географических названий число возможных флективных классов для существительных значительно уменьшается. Это происходит:

во-первых, за счет классов, относящихся к одушевленным существительным как мужского, так и женского рода (таких классов соответственно 19 и 8).

во-вторых, флективные классы для неодушевленных существительных зачастую различаются типом склонения лишь во множественном числе, однако для тех географических названий, которые соответствуют форме единственного числа нет необходимости генерировать словоформы множественного числа.

# Работа с web-приложением автоматизированной генерации словоформ

1.Пользователю предоставляется возможность ввести географическое название.

2.При выборе существительного на следующем шаге необходимо указать род или выбрать к какому типу это слово относится.

3.Далее, для существительного нужно выбрать номер флективного класса, которому соответствует слово.

4. Просклоняв данное слово по указанным формам, и сравнив полученные окончания с окончаниями из таблицы, можно однозначно определить номер флективного класса.

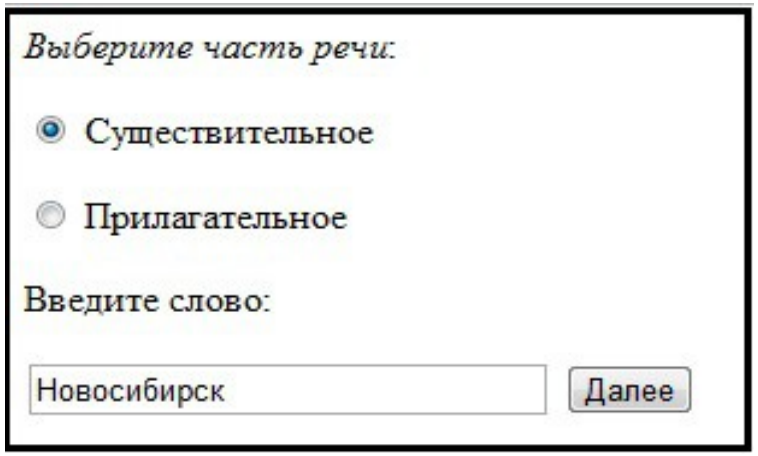



Рисунок 7 – Выбор части речи

При выборе существительного на следующем шаге необходимо указать род или выбрать к какому типу это слово относится.

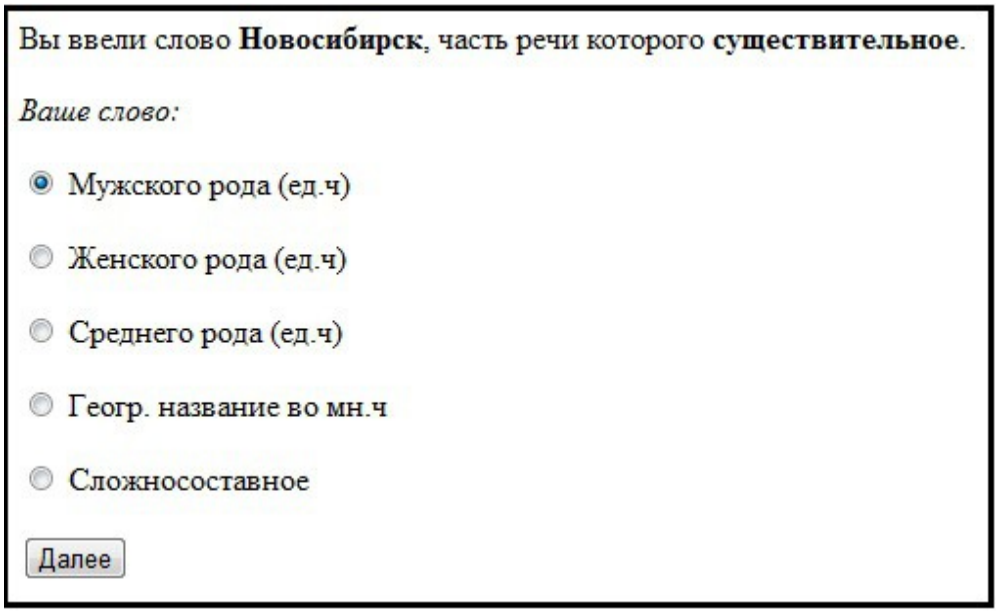



Рисунок 8- Выбор рода слова «Новосибирск»

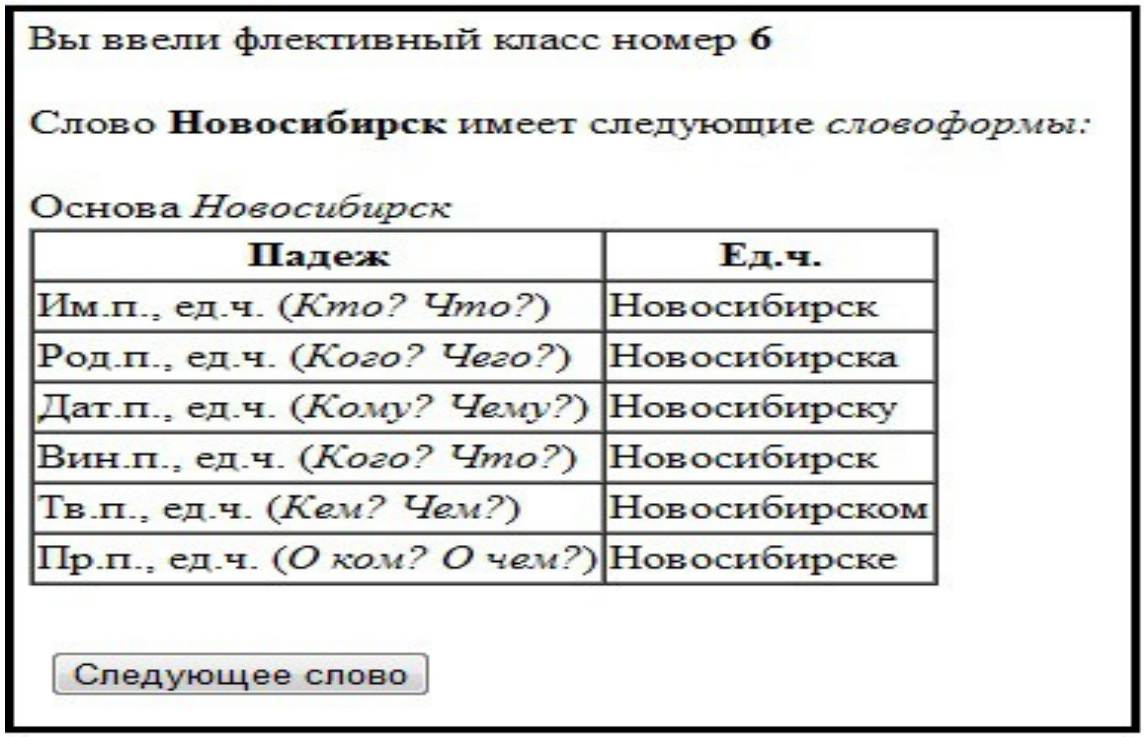
Далее, для существительного нужно выбрать номер флективного класса, которому соответствует слово.

Вы ввели слово **Новосибирск**, часть речи которого **существительное**.

*Выберите номер флективного класса, слово-представитель которого склоняется так же как **Новосибирск**:*

| № класса | Слово-представитель | Им.п., ед.ч. | Тв.п., ед.ч. *Кем? Чем?* |
|---|---|---|---|
| 1 | Курган | + | ом |
| 2 | Воронеж | + | ом |
| 6 | Томск | + | ом |
| 7 | Таганрог | + | ом |
| 10 | Миасс | + | ом |
| 11 | Череповец | + | ем |
| 14 | Энгельс | + | ом |
| 15 | Владикавказ | + | ом |
| 16 | Актаныш | + | ем |
| 17 | Лангепас | + | ом |

Рисунок 9- Выбор рода слова «Новосибирск»



Вы ввели флективный класс номер **6**

Слово **Новосибирск** имеет следующие *словоформы:*

Основа *Новосибирск*

| Падеж | Ед.ч. |
|---|---|
| Им.п., ед.ч. (*Кто? Что?*) | Новосибирск |
| Род.п., ед.ч. (*Кого? Чего?*) | Новосибирска |
| Дат.п., ед.ч. (*Кому? Чему?*) | Новосибирску |
| Вин.п., ед.ч. (*Кого? Что?*) | Новосибирск |
| Тв.п., ед.ч. (*Кем? Чем?*) | Новосибирском |
| Пр.п., ед.ч. (*О ком? О чем?*) | Новосибирске |

Следующее слово

Рисунок 10 - Вывод списка словоформ «Новосибирск»

Программа выводит список словоформ, так как обычно слово может иметь одинаковые окончания в разных формах . При нажатии на кнопку «Следующее слово» приложение возвращается на первый шаг.

Выявление омонимов и и многозначных названий

Географические названия бывают омонимичны другим словам, являющимися именами как нарицательными так и собвственными.

*Город Орёл*
*птица орёл*

***Река Белая*** ***цвет вещи белая***

***Город Киров***
***фамилия Киров***

***Город Кострома*** ***Кострома-Божество весны***

1. Заранее в процессе работы с тезаурусом составить список географических названий, имеющих такие омонимы и «многозначных» названий.

2. Омонимы имен нарицательных выявляются в процессе сравнения тезауруса географических названий с общим лексическим словарем, используемым программной библиотеке.

3. Омонимы имен собственных могут быть обнаружены при сравнении тезауруса с биографическими, мифологическими и т.п. словарями.

# Этапы алгоритма

I. Извлечение из текста документа всех географических названий, входящих в тезаурус.

2. Создание предварительного индекса «номер названия» – «позиция слова в названии» – «название в символьном представлении».

3. Добавление встретившихся неизвестных слов в тезаурус, где им присваиваются идентификационные номера.

4. Переработка индекса в формат «номер названия» – «позиция в тексте» – «номер названия из лексического словаря».

5. Сбор статистики о длинах названий для реализации поиска и идентификации составных названий (т.е. названий, состоящих более чем из одного слова).

6. Сбор статистики о количестве вхождений отдельных слов для оптимизации поиска путем исключения из рассмотрения терминов, заведомо отсутствующих в тексте.

подсчет количества вхождений названий в текст (тексты). Ее этапы:

7. Подсчет возможных комбинаций «текст» – «название», основанный на статистике вхождения отдельных слов.

8. Нахождение всех потенциально возможных мест вхождения каждого названия в текст (тексты) на основе наличия хотя бы одного общего слова из лексического словаря. Позиция каждого потенциально возможного вхождения фиксируется.

Из всей структуры записи нас будут интересовать несколько полей, а именно:
-заголовок статьи
-аннотация к статье
-источник
-место публикации
Заголовок статьи и аннотацию отнесем к контенту, источник и место публикации – к контексту.

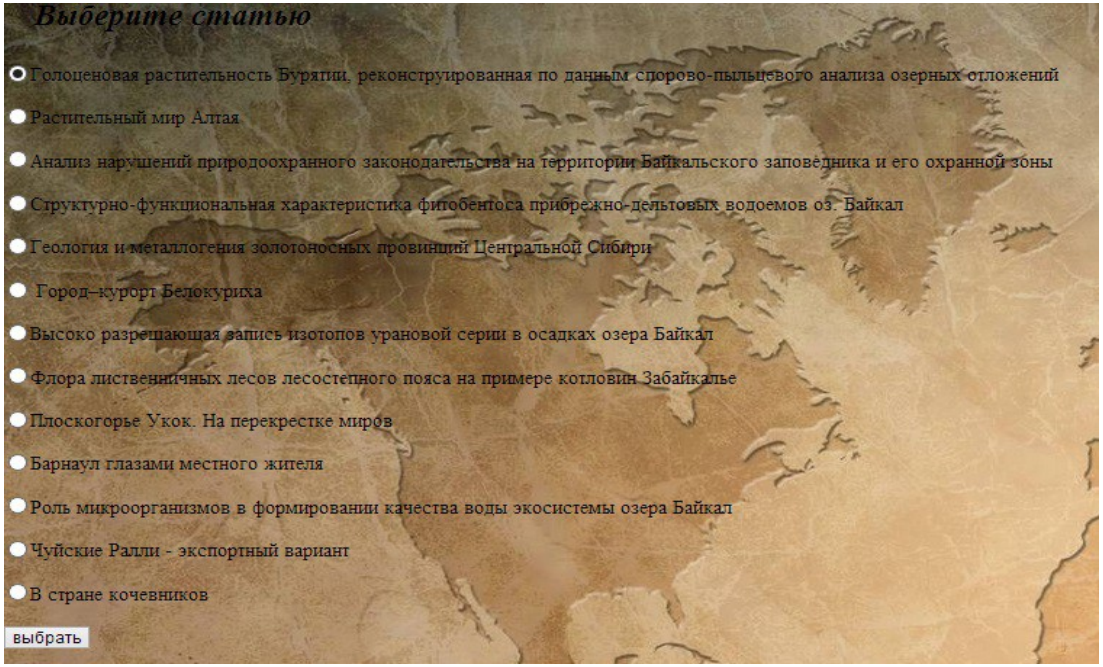


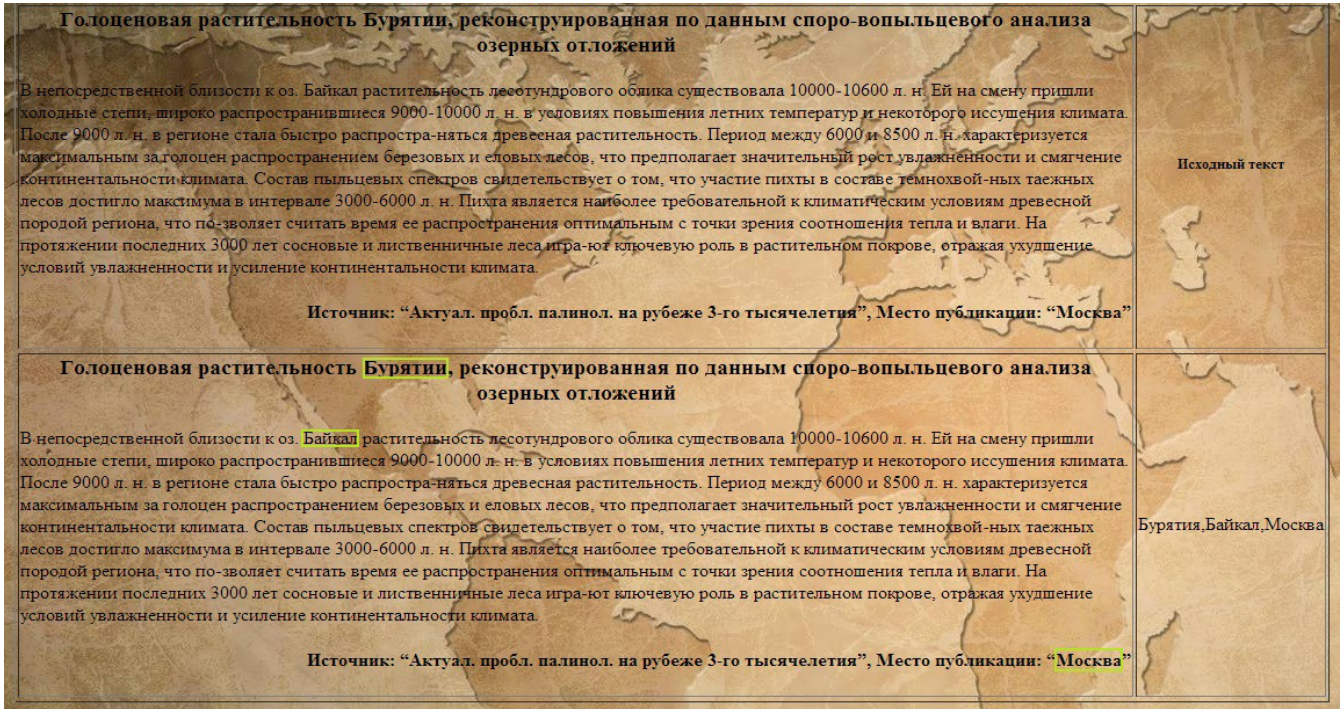

Рисунок 11 – Выбор статьи №1

**1).** Заголовок: “Голоценовая растительность Бурятии, реконструированная по данным спорово-пыльцевого анализа озерных отложений”.

Источник: “Актуальные проблемы палинол. на рубеже 3-го тысячелетия”

Место публикации: “Москва”.

Аннотация: “В непосредственной близости к оз. Байкал растительность лесотундрового облика существовала 10000-10600 л. н. Ей на смену пришли холодные степи, широко распространившиеся 9000-10000 л. н. в условиях повышения летних температур и некоторого иссушения климата. После 9000 л. н. в регионе стала быстро распространяться древесная растительность. Период между 6000 и 8500 л. н. характеризуется максимальным за голоцен распространением березовых и еловых лесов, что предполагает значительный рост увлажненности и смягчение континентальности климата. Состав пыльцевых спектров свидетельствует о том, что участие пихты в составе темнохвойных таежных лесов достигло максимума в интервале 3000-6000 л. н. Пихта является наиболее требовательной к климатическим условиям древесной породой региона, что позволяет считать время ее распространения оптимальным с точки зрения соотношения тепла и влаги. На протяжении последних 3000 лет сосновые и лиственничные леса играют ключевую роль в растительном покрове, отражая ухудшение условий увлажненности и усиление континентальности климата.”



| | |
|---|---|
| **Голоценовая растительность Бурятии, реконструированная по данным споро-вопыльцевого анализа озерных отложений**<br><br>В непосредственной близости к оз. Байкал растительность лесотундрового облика существовала 10000-10600 л. н. Ей на смену пришли холодные степи, широко распространившиеся 9000-10000 л. н. в условиях повышения летних температур и некоторого иссушения климата. После 9000 л. н. в регионе стала быстро распростра-няться древесная растительность. Период между 6000 и 8500 л. н. характеризуется максимальным за голоцен распространением березовых и еловых лесов, что предполагает значительный рост увлажненности и смягчение континентальности климата. Состав пыльцевых спектров свидетельствует о том, что участие пихты в составе темнохвой-ных таежных лесов достигло максимума в интервале 3000-6000 л. н. Пихта является наиболее требовательной к климатическим условиям древесной породой региона, что по-зволяет считать время ее распространения оптимальным с точки зрения соотношения тепла и влаги. На протяжении последних 3000 лет сосновые и лиственничные леса игра-ют ключевую роль в растительном покрове, отражая ухудшение условий увлажненности и усиление континентальности климата.<br><br>**Источник: “Актуал. пробл. палинол. на рубеже 3-го тысячелетия”, Место публикации: “Москва”** | **Исходный текст** |
| **Голоценовая растительность Бурятии, реконструированная по данным споро-вопыльцевого анализа озерных отложений**<br><br>В непосредственной близости к оз. Байкал растительность лесотундрового облика существовала 10000-10600 л. н. Ей на смену пришли холодные степи, широко распространившиеся 9000-10000 л. н. в условиях повышения летних температур и некоторого иссушения климата. После 9000 л. н. в регионе стала быстро распростра-няться древесная растительность. Период между 6000 и 8500 л. н. характеризуется максимальным за голоцен распространением березовых и еловых лесов, что предполагает значительный рост увлажненности и смягчение континентальности климата. Состав пыльцевых спектров свидетельствует о том, что участие пихты в составе темнохвой-ных таежных лесов достигло максимума в интервале 3000-6000 л. н. Пихта является наиболее требовательной к климатическим условиям древесной породой региона, что по-зволяет считать время ее распространения оптимальным с точки зрения соотношения тепла и влаги. На протяжении последних 3000 лет сосновые и лиственничные леса игра-ют ключевую роль в растительном покрове, отражая ухудшение условий увлажненности и усиление континентальности климата.<br><br>**Источник: “Актуал. пробл. палинол. на рубеже 3-го тысячелетия”, Место публикации: “Москва”** | Бурятия,Байкал,Москва |

Рисунок 12 – Результаты обработки ст. №1

Названия, относящиеся к контенту (заголовок, аннотация): Бурятия, Байкал.
Названия, относящиеся к контексту (источник, место публикации): Москва.

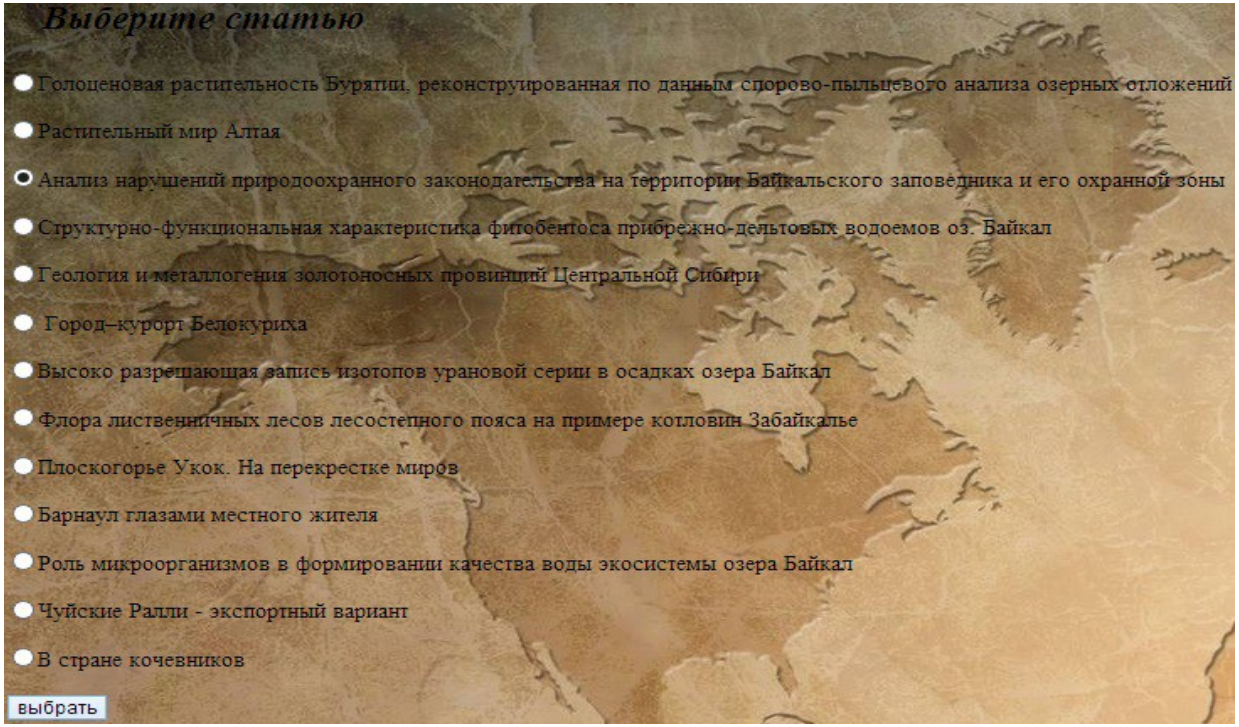
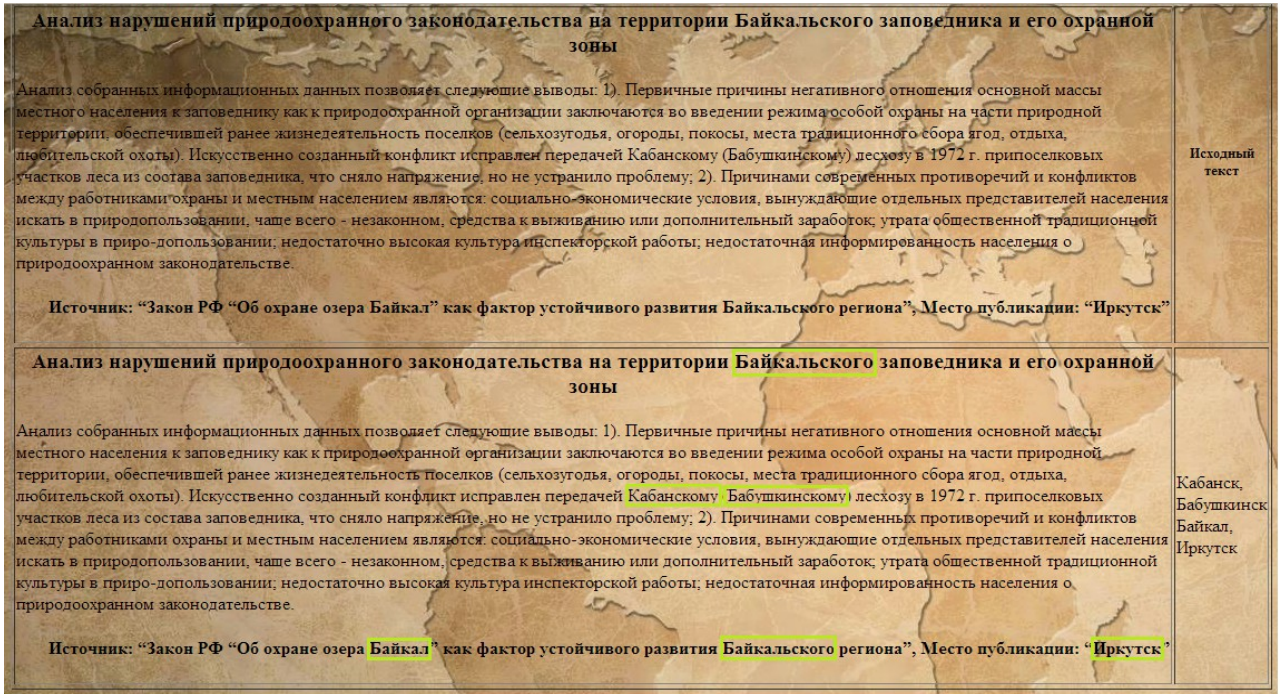



Рисунок 13 – Выбор статьи №2

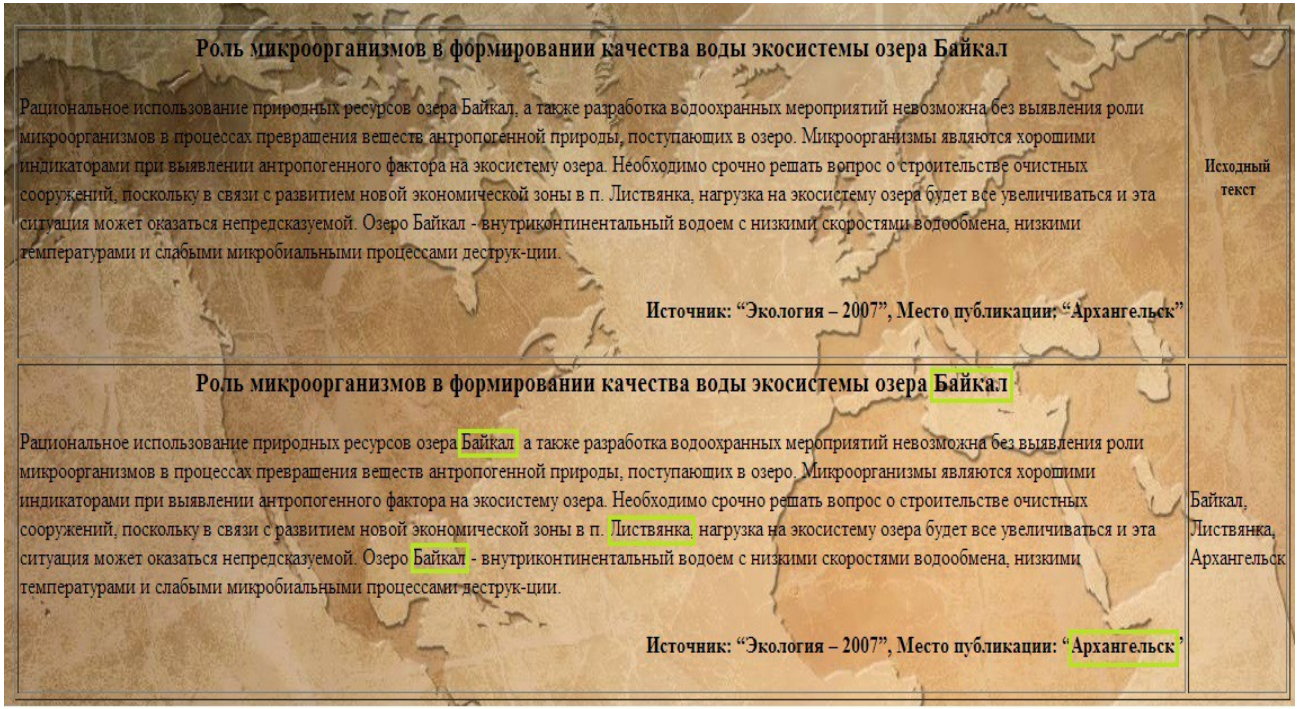


| | |
|---|---|
| **Анализ нарушений природоохранного законодательства на территории Байкальского заповедника и его охранной зоны**<br><br>Анализ собранных информационных данных позволяет следующие выводы: 1). Первичные причины негативного отношения основной массы местного населения к заповеднику как к природоохранной организации заключаются во введении режима особой охраны на части природной территории, обеспечившей ранее жизнедеятельность поселков (сельхозугодья, огороды, покосы, места традиционного сбора ягод, отдыха, любительской охоты). Искусственно созданный конфликт исправлен передачей Кабанскому (Бабушкинскому) лесхозу в 1972 г. припоселковых участков леса из соотава заповедника, что сняло напряжение, но не устранило проблему; 2). Причинами современных противоречий и конфликтов между работниками охраны и местным населением являются: социально-экономические условия, вынуждающие отдельных представителей населения искать в природопользовании, чаще всего - незаконном, средства к выживанию или дополнительный заработок; утрата общественной традиционной культуры в приро-допользовании; недостаточно высокая культура инспекторской работы; недостаточная информированность населения о природоохранном законодательстве.<br><br>**Источник: "Закон РФ "Об охране озера Байкал" как фактор устойчивого развития Байкальского региона", Место публикации: "Иркутск"** | **Исходный текст** |
| **Анализ нарушений природоохранного законодательства на территории Байкальского заповедника и его охранной зоны**<br><br>Анализ собранных информационных данных позволяет следующие выводы: 1). Первичные причины негативного отношения основной массы местного населения к заповеднику как к природоохранной организации заключаются во введении режима особой охраны на части природной территории, обеспечившей ранее жизнедеятельность поселков (сельхозугодья, огороды, покосы, места традиционного сбора ягод, отдыха, любительской охоты). Искусственно созданный конфликт исправлен передачей Кабанскому (Бабушкинскому) лесхозу в 1972 г. припоселковых участков леса из состава заповедника, что сняло напряжение, но не устранило проблему; 2). Причинами современных противоречий и конфликтов между работниками охраны и местным населением являются: социально-экономические условия, вынуждающие отдельных представителей населения искать в природопользовании, чаще всего - незаконном, средства к выживанию или дополнительный заработок; утрата общественной традиционной культуры в приро-допользовании; недостаточно высокая культура инспекторской работы; недостаточная информированность населения о природоохранном законодательстве.<br><br>**Источник: "Закон РФ "Об охране озера Байкал" как фактор устойчивого развития Байкальского региона", Место публикации: "Иркутск"** | Кабанск, Бабушкинск Байкал, Иркутск |

Рисунок 14– Результаты обработки статьи №2.

Названия, относящиеся к контенту (заголовок, аннотация): Байкал, Кабанск, Бабушкинск.

Названия, относящиеся к контексту (источник, место публикации): Байкал, Иркутск.



| | |
|---|---|
| **Роль микроорганизмов в формировании качества воды экосистемы озера Байкал**<br><br>Рациональное использование природных ресурсов озера Байкал, а также разработка водоохранных мероприятий невозможна без выявления роли микроорганизмов в процессах превращения веществ антропогенной природы, поступающих в озеро. Микроорганизмы являются хорошими индикаторами при выявлении антропогенного фактора на экосистему озера. Необходимо срочно решать вопрос о строительстве очистных сооружений, поскольку в связи с развитием новой экономической зоны в п. Листвянка, нагрузка на экосистему озера будет все увеличиваться и эта ситуация может оказаться непредсказуемой. Озеро Байкал - внутриконтинентальный водоем с низкими скоростями водообмена, низкими температурами и слабыми микробиальными процессами деструк-ции.<br><br>**Источник: "Экология – 2007", Место публикации: "Архангельск"** | **Исходный текст** |
| **Роль микроорганизмов в формировании качества воды экосистемы озера Байкал**<br><br>Рациональное использование природных ресурсов озера Байкал, а также разработка водоохранных мероприятий невозможна без выявления роли микроорганизмов в процессах превращения веществ антропогенной природы, поступающих в озеро. Микроорганизмы являются хорошими индикаторами при выявлении антропогенного фактора на экосистему озера. Необходимо срочно решать вопрос о строительстве очистных сооружений, поскольку в связи с развитием новой экономической зоны в п. Листвянка, нагрузка на экосистему озера будет все увеличиваться и эта ситуация может оказаться непредсказуемой. Озеро Байкал - внутриконтинентальный водоем с низкими скоростями водообмена, низкими температурами и слабыми микробиальными процессами деструк-ции.<br><br>**Источник: "Экология – 2007", Место публикации: "Архангельск"** | Байкал, Листвянка, Архангельск |

Рисунок 15 – Результаты обработки статьи №3.

Названия, относящиеся к контенту (заголовок, аннотация): Байкал, Листвянка.

Названия, относящиеся к контексту (источник, место публикации): Архангельск.

## ЗАКЛЮЧЕНИЕ

Из полученных результатов можно сделать следующие выводы: алгоритм извлекает географические названия с достаточной точностью, кроме сокращенных, омономичных названий по причине сложностей с их определением.

В целом, работоспособность общего алгоритма была проверена при решении ряда практических задач. Работа может быть использована в задачах, где требуется поиск и извлечение ключевых слов с текстов, с применением тезауруса предметной области, а также в геоинформационных системах.

Спасибо за внимание!