**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«Тихоокеанский государственный университет»**

Высшая школа Кибернетики и цифровых технологий

Создание системы «База Знаний ТОГУ» с полнотекстовым поиском на основе лексем естественного языка

Текстовый документ МД

МД.220000598.ТД

Раздаточный материал

Руководитель доцент к.ф.-м.н. Вихтенко Э.М.

Студент гр. ПИИ(м)-31 Забавин А.С.

Хабаровск – 2025 г.

Содержание

Введение 5

1 Анализ предметной области 9

1.1 Основные понятия 9

1.2 Задача обработки естественного языка в компьютерной лингвистике 9

1.3 Полнотекстовый поиск 12

1.4 Существующие решения полнотекстового поиска 13

1.5 Метрики релевантности. Оценка качества ранжирования в задаче поиска 16

1.6 Существующие модели для NLP обработки текстов 19

1.7 Описание концепции «словесных вложений» рассматриваемых моделей, Векторные представления слов 29

1.8 Подведение итогов 36

2 Описание разработанных алгоритмов 39

2.1 Алгоритм синтаксического анализа запроса, выявление основной части запроса 39

2.2 Алгоритм оптимизации по семантической близости и TF-IDF 41

3 Описание разработанной программы 44

3.1 Назначение программы и выбранный инструментарий 44

3.2 Описание интерфейса программы 46

3.3 Тестирование качества алгоритма оптимизации запроса 55

3.4 Описание технической реализации программы 59

Заключение 62

Список использованных источников 63

Приложение А – Исходный код 67

Приложение Б – антиплагиат 83

**Цели и задачи**

Цель данного исследования состоит в разработке программного средства, использующего полнотекстовый поиск Postgres и RNN искусственную нейронную сеть семейств Word2vec и GloVe для оптимизации запросов к базе данных.

Задачи состоят в изучении качества поиска в различных вариациях - при простом поиске по вхождению текста, при индексировании на основе «частотной важности» слов в документе, а также используя оптимизации поискового запроса на основе семантической близости и синтаксической важности членов предложения поискового запроса. А также разработке структуры хранения знаний в базе, их индексации, а также разработка алгоритмов оптимизации поискового запроса на основе исследованных данных.

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд2.TIF

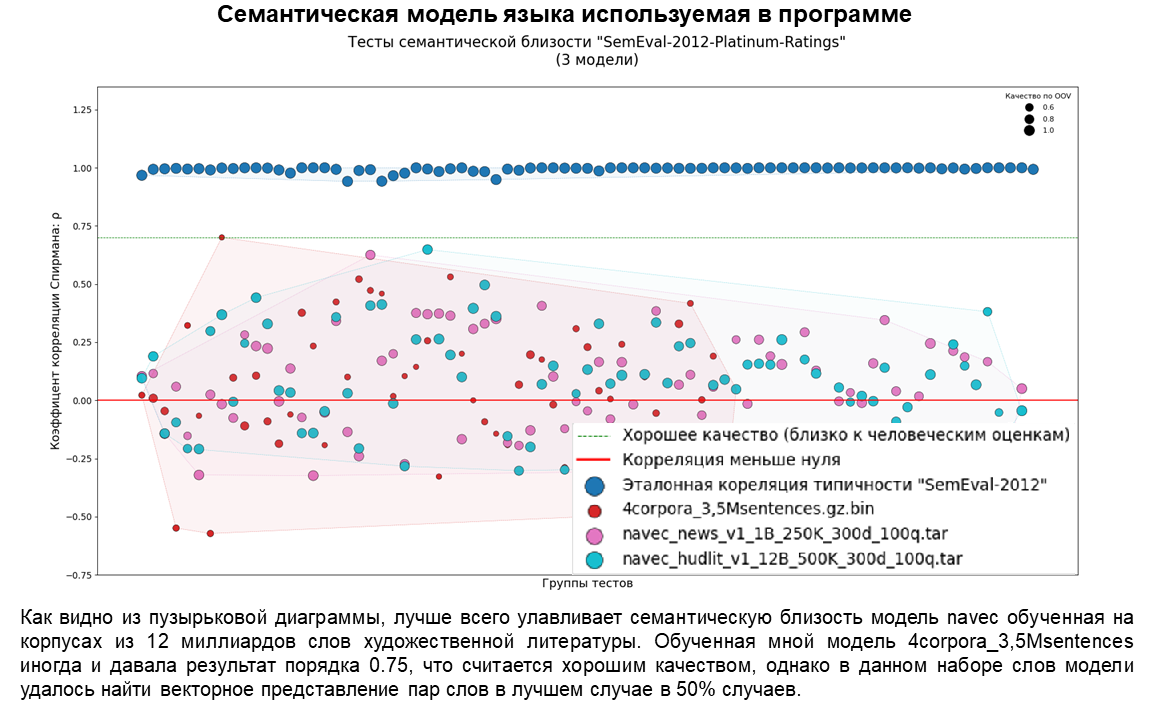
M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд4.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд9.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд10.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд11.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд12.TIF



M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд14.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд15.TIF

**Проведенные тесты алгоритмов оптимизации**

Было задано 10 запросов для теста. Вопрос строково идентичен вопросу в базе у двух запросов «Как назывался свод законов в Древней Руси?» и «Какого цвета язык у жирафа?». Убраны/изменены союзы, предлоги, окончания вопроса присутствующего в базе знаний у 2 запросов «Какова длинна кровеносных сосудов человека?» (убрано обобщение «всех»), «Сколько пузырьков углекислого газа содержит шампанское?» (изменены окончания и убран предлог «в»). Остальные 6 запросов перефразированы с использованием синонима или изменен порядок членов предложения.

Таблица 1 – Результаты точности поиска p@K и ap@K для поиска по вхождению строки, при K=10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Запрос | Точное соответствие | Изменение | Перефразирование | p@K | ap@K |
| 1 | Какова длинна кровеносных сосудов человека? |  | + |  | 0.0 | 0.0 |
| 2 | Сколько в человеке кровеносных сосудов? |  |  | + | 0.0 | 0.0 |
| 3 | Какого цвета язык у жирафа? | + |  |  | 1.0 | 1.0 |
| 4 | Кто проживает на дне моря? |  |  | + | 0.0 | 0.0 |
| 5 | Кто живет на дне? |  |  | + | 0.0 | 0.0 |
| 6 | Сколько пузырьков углекислого газа содержит шампанское? |  | + |  | 0.0 | 0.0 |
| 7 | Как получить документ с оценками для иностранного вуза? |  |  | + | 0.0 | 0.0 |
| 8 | Не появились индивидуальные достижения в анкете абитуриента |  |  | + | 0.0 | 0.0 |
| 9 | Я преподаватель кафедры и не могу зайти на сайт и заполнить журнал преподавателя |  |  | + | 0.0 | 0.0 |
| 10 | Как назывался свод законов в Древней Руси? | + |  |  | 1.0 | 1.0 |

Средневзвешенная точность на K элементов для данного алгоритма поиска ответов по вхождению строки представлена на формуле :

|  |  |
| --- | --- |
|  | (1) |

Таблица 2 – Результаты точности поиска p@K и ap@K для полнотекстового поиска Postgres, при K=10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Запрос | Точное соответствие | Изменение | Перефразирование | p@K | ap@K |
| 1 | Какова длинна кровеносных сосудов человека? |  | + |  | 1.0 | 1.0 |
| 2 | Сколько в человеке кровеносных сосудов? |  |  | + | 0.0 | 0.0 |
| 3 | Какого цвета язык у жирафа? | + |  |  | 1.0 | 1.0 |
| 4 | Кто проживает на дне моря? |  |  | + | 0.0 | 0.0 |
| 5 | Кто живет на дне? |  |  | + | 0.0 | 0.0 |
| 6 | Сколько пузырьков углекислого газа содержит шампанское? |  | + |  | 1.0 | 1.0 |
| 7 | Как получить документ с оценками для иностранного вуза? |  |  | + | 0.0 | 0.0 |
| 8 | Не появились индивидуальные достижения в анкете абитуриента |  |  | + | 0.0 | 0.0 |
| 9 | Я преподаватель кафедры и не могу зайти на сайт и заполнить журнал преподавателя |  |  | + | 0.0 | 0.0 |
| 10 | Как назывался свод законов в Древней Руси? | + |  |  | 1.0 | 1.0 |

|  |  |
| --- | --- |
|  | (2) |

Таблица 3 – Результаты точности поиска p@K и ap@K для полнотекстового поиска с NLP оптимизацией, при K=10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Запрос | Точное соответствие | Изменение | Перефразирование | p@K | ap@K |
| 1 | Какова длинна кровеносных сосудов человека? |  | + |  | 1.0 | 1.0 |
| 2 | Сколько в человеке кровеносных сосудов? |  |  | + | 1.0 | 1.0 |
| 3 | Какого цвета язык у жирафа? | + |  |  | 1.0 | 1.0 |
| 4 | Кто проживает на дне моря? |  |  | + | 1.0 | 1.0 |
| 5 | Кто живет на дне? |  |  | + | 1.0 | 1.0 |
| 6 | Сколько пузырьков углекислого газа содержит шампанское? |  | + |  | 1.0 | 1.0 |
| 7 | Как получить документ с оценками для иностранного вуза? |  |  | + | 0.0 | 0.0 |
| 8 | Не появились индивидуальные достижения в анкете абитуриента |  |  | + | 1.0 | 1.0 |
| 9 | Я преподаватель кафедры и не могу зайти на сайт и заполнить журнал преподавателя |  |  | + | 0.33 | 0.61 |
| 10 | Как назывался свод законов в Древней Руси? | + |  |  | 1.0 | 1.0 |

|  |  |
| --- | --- |
|  | (3) |

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд16.TIF

M:\home\git\magistry2022\study.mag_diploma\Мой диплом\ВКР-набор-документов\Презентации в картинках\Презентация-Забавин v01\Слайд17.TIF