ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

РАЗРАБОТКА ПРОГРАММЫ «БАЗЫ ЗНАНИЙ ТОГУ» С ИСПОЛЬЗОВАНИЕМ ПОЛНОТЕКСТОВОГО ПОИСКА С ПОМОЩЬЮ ЛЕКСЕМ ЕСТЕСТВЕННОГО ЯЗЫКА

Забавин Алексей Сергеевич
Тихоокеанский Государственный Университет
Г. Хабаровск
2025

Предмет работы

Предметом исследования является разработка вопросно-ответной системы базы знаний ТОГУ. Изучение качества поиска — при простом поиске по вхождению текста, при индексировании на основе «частотной важности» слов в документе и полнотекстовом поиске по нему.

А также использование оптимизаций поискового запроса на основе семантической близости и синтаксической важности членов предложения в тексте документа

Объектами исследования являются:

1. хранение информации для QA-системы в базе данных позволяющее решать задачу полнотекстового поиска в ней;

2. частотный алгоритм ранжирования результатов поиска в коллекции документов;

3. методы лексического, синтаксического и семантического анализа текста.

Задача полнотекстового поиска

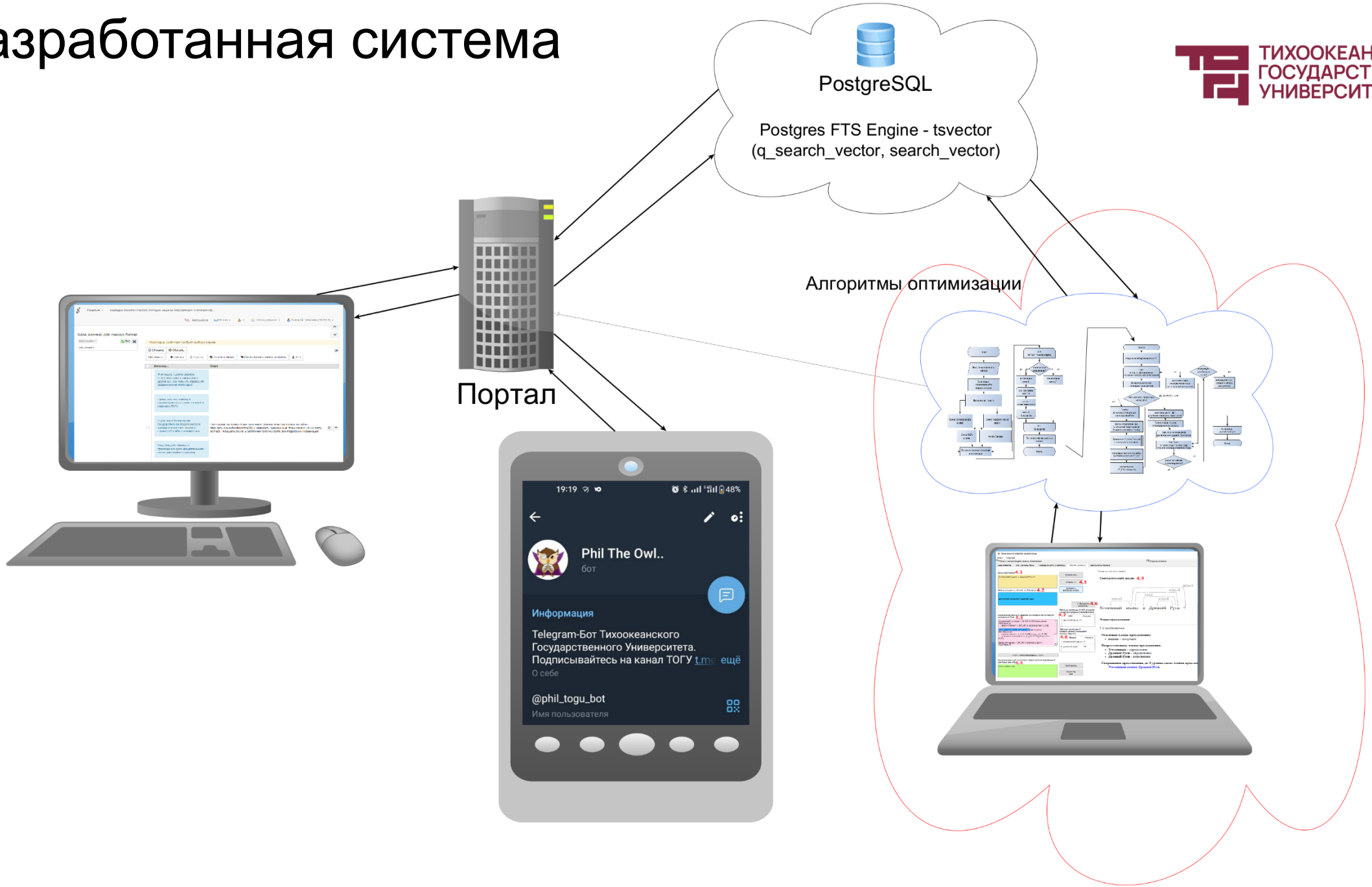
Полнотекстовый поиск предназначен для поиска и ранжирования текстовых данных на основе ключевых слов или фраз, встречающихся в текстовых полях базы данных где стандартные механизмы вроде оператора LIKE недостаточны.

Поиск должен учитывать различные формы слов.

Важным аспектом является не только нахождение документов, но и их сортировка по релевантности. Стандартные SQL-запросы не обладают встроенной поддержкой ранжирования результатов на основе того, насколько близки слова запроса к друг другу в документе или как часто они встречаются.

Разработанная система

ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ



Введение в NLP. Обратная частота встречаемости TF-IDF

ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

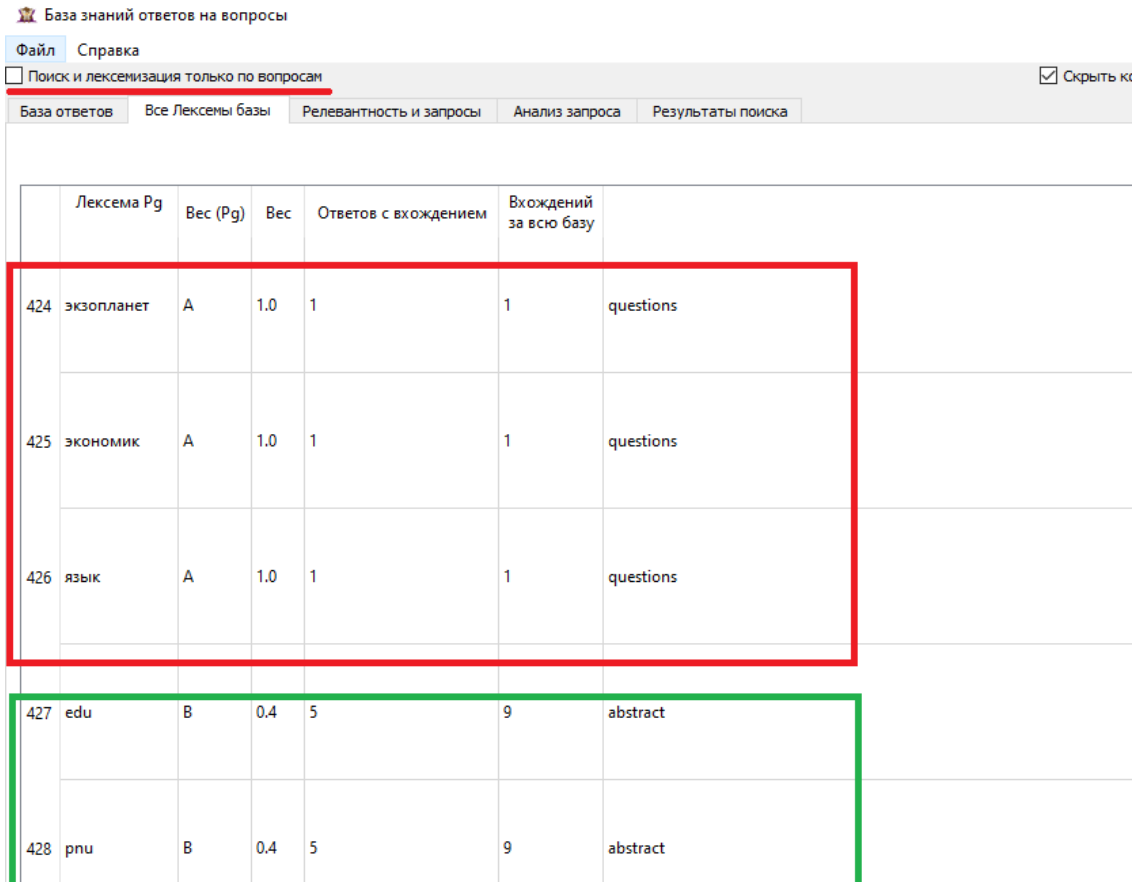
В работе используется движок полнотекстового поиска Postgres. Документы с помощью него индексируются по «лексемам» – базовым синтаксическим единицам представляющим неизменяемые части слов.

Результаты поиска ранжируются в соответствии с статистикой встречаемости слов во всей базе и в документе:

$$TF-IDF(w,d,C)=\frac{count(w,d)}{count(d)} * \log \frac{|C|}{\sum_{d \in C} countif(d,d,w \in d)}$$

Это можно назвать «важностью слова»

Пример разбиения в базе данных:



	Лексема Pg	Вес (Pg)	Вес	Ответов с вхождением	Вхождений за всю базу	
424	экзопланет	A	1.0	1	1	questions
425	экономик	A	1.0	1	1	questions
426	язык	A	1.0	1	1	questions
427	edu	B	0.4	5	9	abstract
428	pnu	B	0.4	5	9	abstract

При добавлении данных автоматически

применяется операция «стемминг» к документу, и строится подобный индекс с подсчетом вхождения лексемы.

Обработка естественного языка

Базовая машина полнотекстового поиска работает лучше стандартного поиска, однако не всегда достаточна для пользовательских запросов на естественном языке.

Чтобы повысить качество поиска нам необходимо углубится в теорию работы с естественным языком (NLP – Natural Langueage Processing)

Основы NLP анализа текста. Эмбеддинги

ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В широком смысле, **эмбеддинг** - это процесс преобразования каких-либо данных (чаще всего текста, но могут быть и изображения, звуки и т.д.) в набор чисел, **векторы**, которые машина может не только хранить, но и с которыми она может работать.

Именно преобразовав слово в числовой вид можно применить аппарат математики и вычислительной техники к NLP-анализу текста

Категория	тип	описание
Текстовые эмбеддинги	Word Embeddings	Эти эмбеддинги преобразуют слова в векторы, так что слова с похожим значением имеют похожие векторные представления
	Sentence Embeddings	Здесь уже идет дело о целых предложениях. Подобные модели создают векторные представления для целых предложений или даже абзацев, улавливая гораздо более тонкие нюансы языка.
Эмбеддинги изображений	CNN	CNN позволяет преобразовать изображения в векторы, которые затем используются для различных задач, например, классификации изображений или даже генерации новых изображений.
	Autoencoders	Автоэнкодеры могут сжимать изображения в более мелкие, плотные векторные представления, которые затем могут быть использованы для различных целей, включая декомпрессию или даже обнаружение аномалий.
Эмбеддинги для других типов данных	Graph Embeddings	Применяются для работы с графовыми структурами (к примеру рекомендательные системы). Это способ представить узлы и связи графа в виде векторов.
	Sequence Embeddings	Используются для анализа последовательностей, например, во временных рядах или в музыке.

Векторные пространства — это математические структуры, состоящие из векторов. Векторы можно понимать как точки в некотором пространстве, которые обладают направлением и величиной. В эмбеддингах, каждый вектор представляет собой уникальное представление объекта, преобразованное в числовую форму.

Размерность вектора определяет, сколько координат используется для описания каждого вектора в пространстве. В эмбеддингах высокая размерность может означать более детализированное представление данных. Векторное пространство для текстовых эмбеддингов может иметь тысячи измерений.

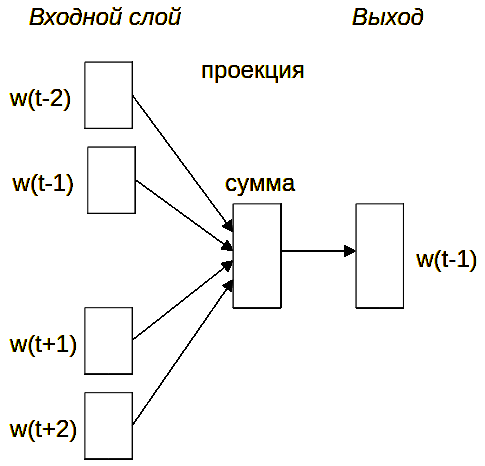
Расстояние между векторами в эмбеддингах измеряется с помощью метрик, таких как *Евклидово расстояние* или *косинусное сходство*. Метрики позволяют оценить, насколько близко или далеко друг от друга находятся различные объекты в векторном пространстве, что является основой для многих алгоритмов машинного обучения, таких как классификация

Используемые технологии. Word2Vec

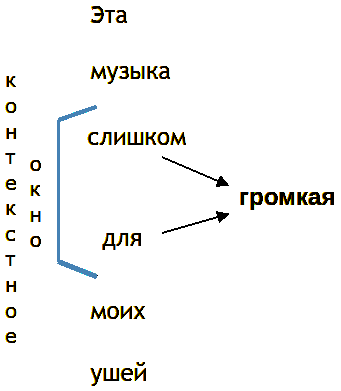
Word2Vec использует нейронные сети для обучения векторных представлений слов из больших наборов текстовых данных. Существуют две основные архитектуры Word2Vec:

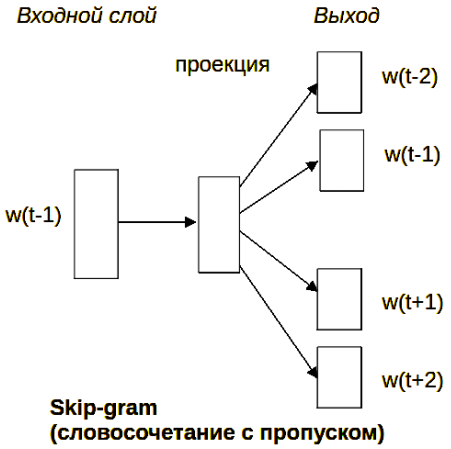
CBOW: предсказывает текущее слово на основе контекста (окружающих слов). Например, в предложении "Собака лает на ___", CBOW попытается угадать недостающее слово (например, "почтальона") на основе окружающих слов.

Skip-gram: работает наоборот по сравнению с CBOW. Использует текущее слово для предсказания окружающих его слов в предложении. Например, если взять слово "кошка", модель попытается предсказать слова, которые часто встречаются в окружении слова "кошка", такие как "мышь", "мяукает" и т.д.

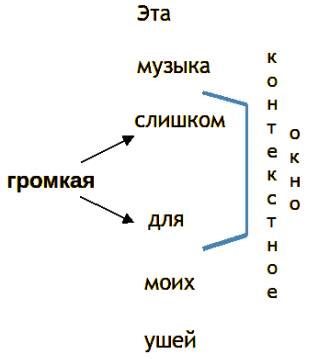


Непрерывный мешок слов





Skip-gram (словосочетание с пропуском)



Ты не должен создавать машину по подобию разума человека

Скользящее окно перемещается по тексту

ты	не	должен	создавать	машину	по	подобию	разума	человека	...
ты	не	должен	создавать	машину	по	подобию	разума	человека	...
ты	не	должен	создавать	машину	по	подобию	разума	человека	...
ты	не	должен	создавать	машину	по	подобию	разума	человека	...
ты	не	должен	создавать	машину	по	подобию	разума	человека	...

Датасет

вход1	вход2	выход
не	должен	создавать
должен	создавать	машину
создавать	машину	по
машину	по	подобию
по	подобию	разума

предыдущий контекст

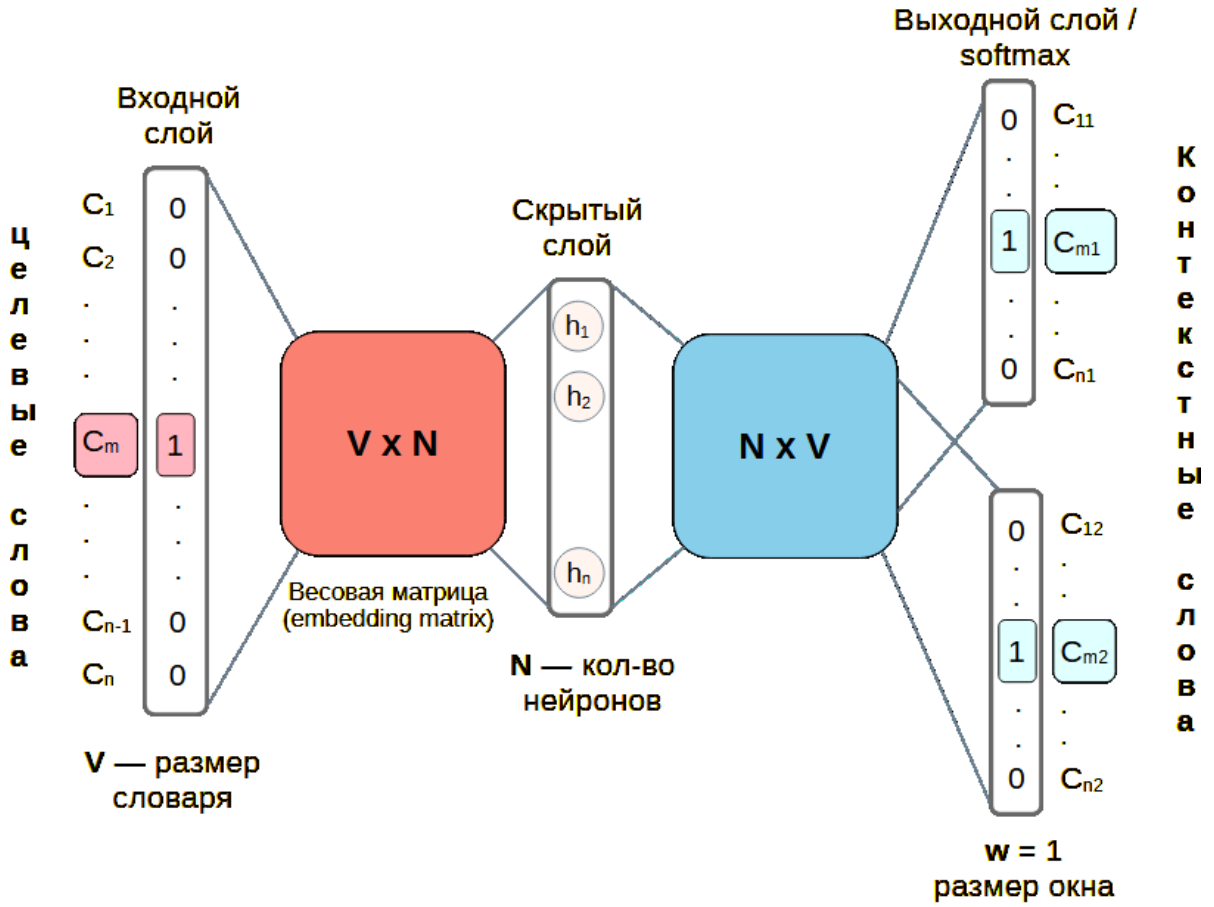
Архитектура модели	Набор тестов на семантико-синтаксическую взаимосвязь слов		Связанность слов MSR (тестовый набор[20])
	Семантическая точность, %	Синтаксическая точность,%	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Ты не должен создавать машину по подобию разума человека

ты	не	должен	создавать	машину	по	подобию	разума	человека	...

входное слово	искомое слово
должен	ты
должен	не
должен	создавать
должен	машину

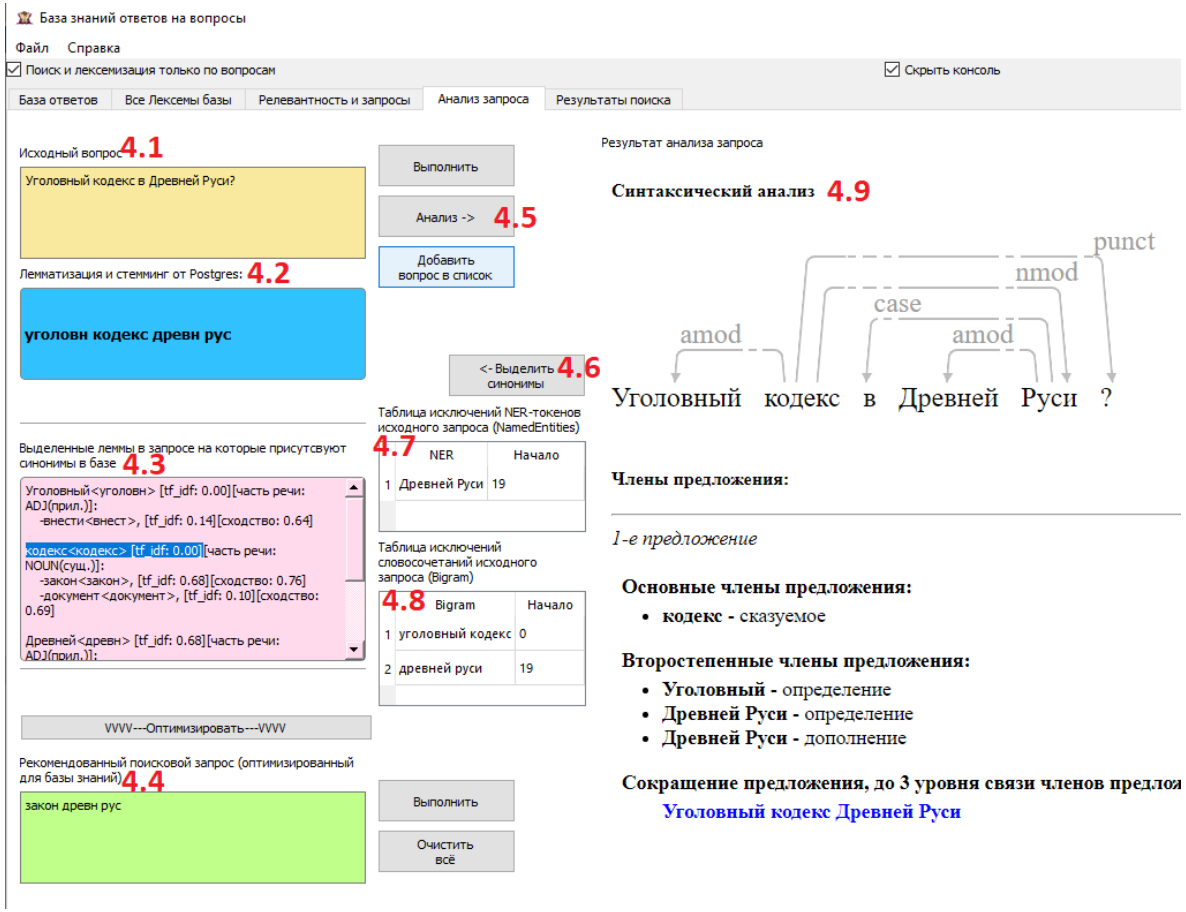
Пример архитектуры Word2vec ИНС (skip-gram), 1 скрытый слой, окно = 1



Объективная (целевая) функция для сети используется для предсказания целевого слова использует логарифмическую сумму вероятностей окружающих n-искомых слов вокруг целевого слова Cm.

V - количество слов в словаре после обучения, каждое слово в словаре описывается как вектор с однократным кодированием (двоичный вектор, в котором только позиция соответствующего слова имеет значение 1), N - количество нейронов (размерность векторного пространства слов). Весовая матрица VxN хранит обученный вектор и моделью предсказываются векторы которые соответствуют словам близким по контексту входному — то есть при обучении находившихся слева и с права в тексте (окно w=1).

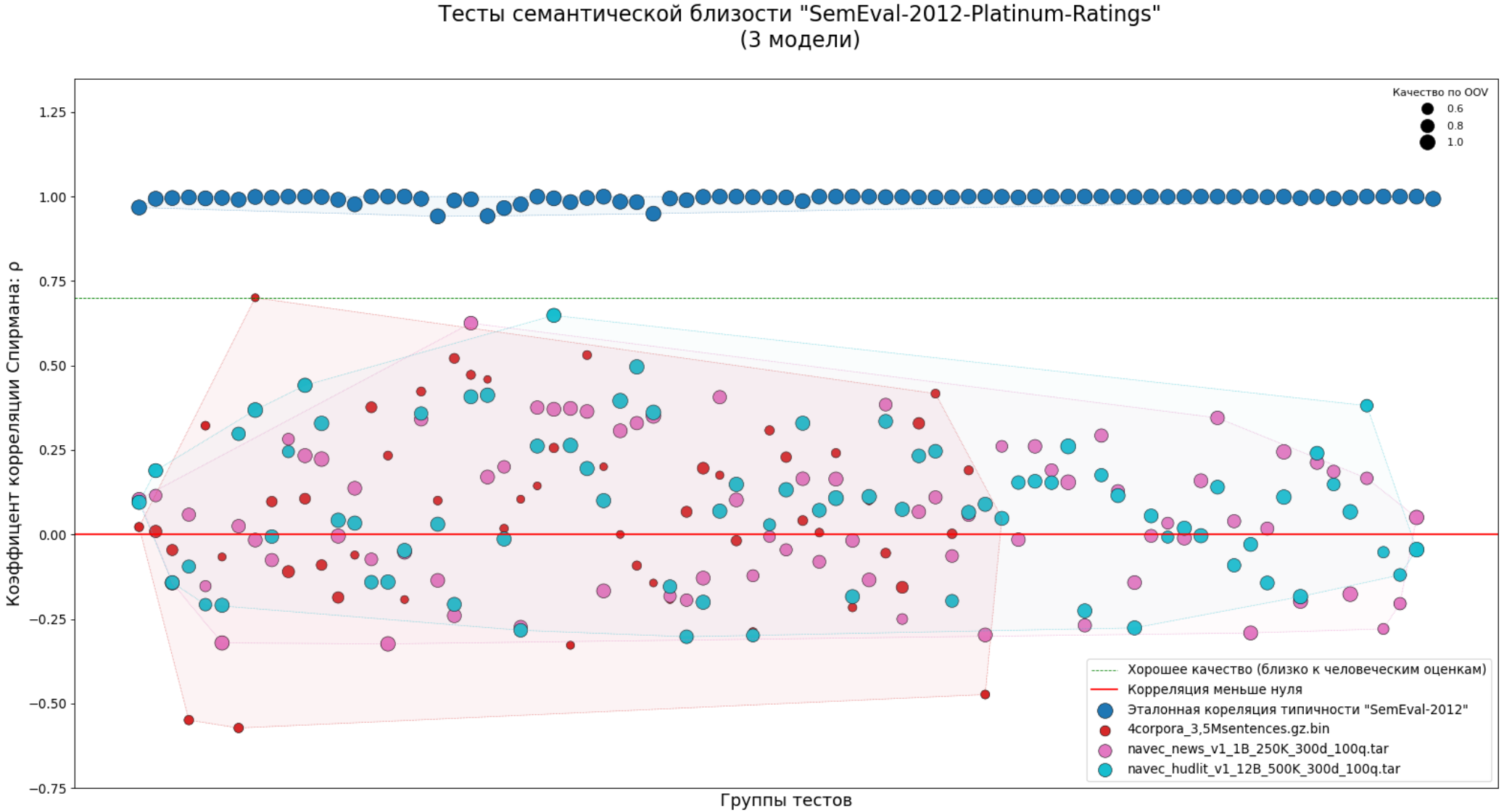
Разработанная программа



Программа использует комплекс из двух алгоритмов: «Алгоритм синтаксического анализа запроса, выявление основной части запроса» и «Алгоритм оптимизации по семантической близости и TF-IDF» для модификации пользовательского запроса к базе данных.

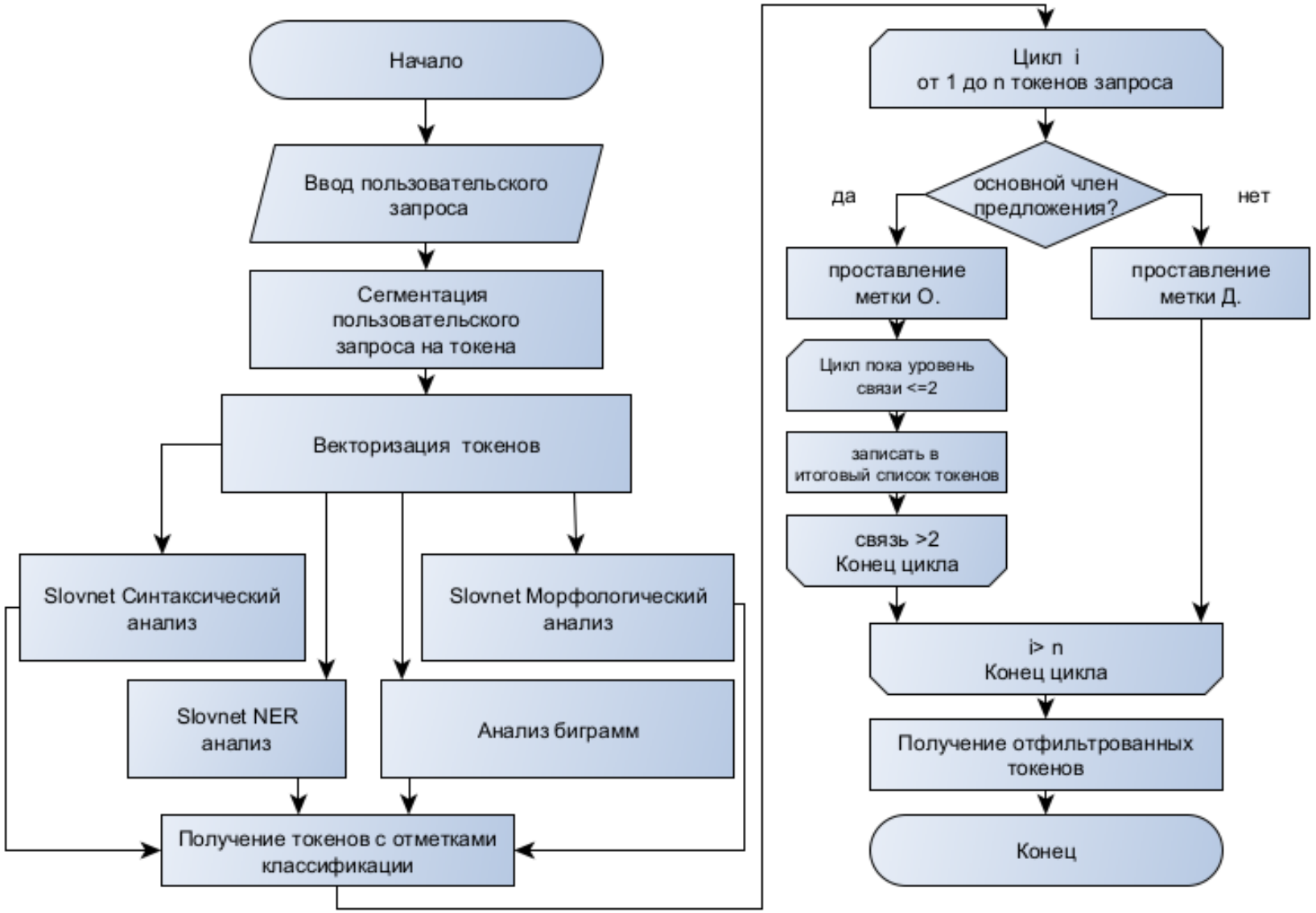
В работе алгоритма используется как уже обученные модели из пакета gensim и natasha, так и полностью самостоятельно обученная фразовая Word2Vec модель словосочетаний

Семантическая модель языка используемая в программе

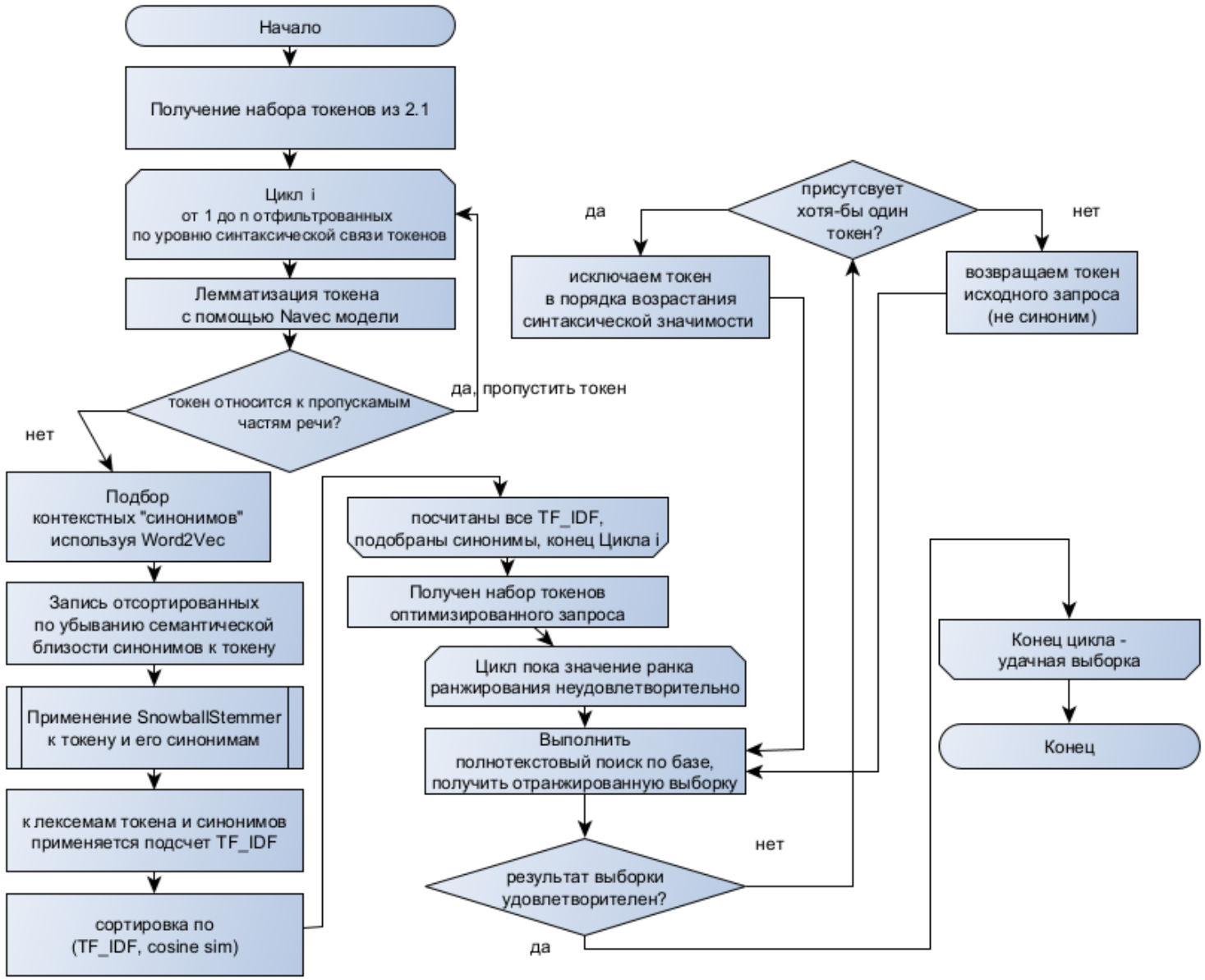


Как видно из пузырьковой диаграммы, лучше всего улавливает семантическую близость модель navec обученная на корпусах из 12 миллиардов слов художественной литературы. Обученная мной модель 4corpora_3,5Msentences иногда и давала результат порядка 0.75, что считается хорошим качеством, однако в данном наборе слов модели удалось найти векторное представление пар слов в лучшем случае в 50% случаев.

«Алгоритм синтаксического анализа запроса, выявление основной части запроса»



«Алгоритм оптимизации по семантической близости и TF-IDF»



ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Таблицы оценки релевантности

(p@K и ap@K, при K=10**)**

Таблица 1 – для поиска по вхождению

№	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

Таблица 2 – для полнотекстового поиска Postgres

№	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

Таблица 3 – для полнотекстового поиска с NLP оптимизацией

№	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

Полученные результаты эффективности

Ранжирование — задача сортировки набора элементов из соображения их релевантности. Чаще всего релевантность понимается по отношению к некому объекту. В задаче информационного поиска объект — это запрос, элементы — всевозможные документы (ссылки на них), а релевантность — соответствие документа запросу.

Для релевантности существует метрика: Средняя точность на k-элементах (map@K)

Были проведены расчеты для 10 поисковых запросов с размеченной релевантностью на базе из 100 вопросов:

Тип поиска	map@K
Поиск по вхождению строки	0,2
Полнотекстовый поиск Postgres	0,4
Полнотекстовый поиск с NLP оптимизацией	0,861

ТИХООКЕАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Вопросы

2025