Студент: Казаева Татьяна, ИУ7-66Б
Научный руководитель: Строганов Юрий Владимирович

Разработка базы данных

для хранения и аналитики результатов статистического опроса

2022 г.

Цель и задачи работы

Цель работы: спроектировать и реализовать программное обеспечение для проведения анкетированного опроса экспертов о тональности прилагательных.

Задачи:

проанализировать варианты хранения данных и выбрать подходящий вариант;

спроектировать базу данных, описать ее сущности и связи;

разработать систему разметки интерфейса.

Задача оценки тональности

Тональность текста *определяется*:

субъектом - автором высказывания;

объектом - то, о чем высказывается субъект;

тональной оценкой – эмоциональным отношением автора к такому объекту.

Тональность текста *используется*:

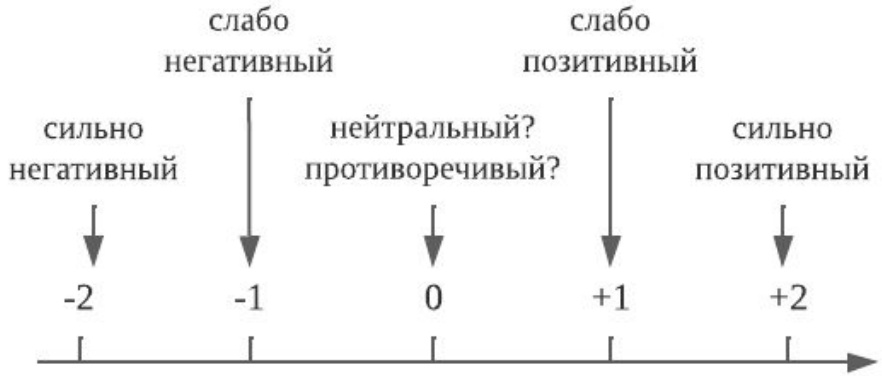
интернет-магазинами, получающими множество отзывов на товары каждый день;

в новостных текстах для выделения мнений об отношениях, третьих лицах;

и.т.д

Общие словари **не могут** рассматриваться как полноценная альтернатива классическим подходам оценки мнений на основе массовых опросов - требуются заранее подготовленные размеченные данные.

Шкалы оценки тональных слов(1/2)

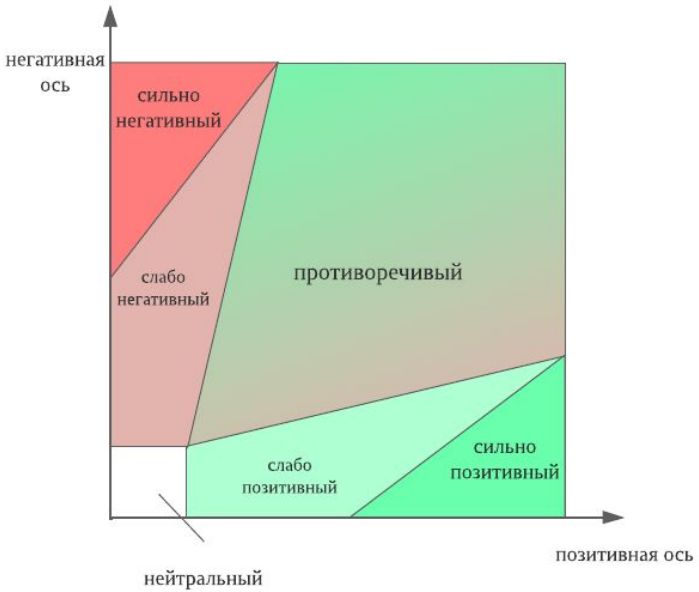


Плоская шкала оценки

Возникает неопределенность:

Нулевое значение - отсутствие эмотивного окраса, так и присутствие позитивного и негативного окраса в равной мере.

Шкалы оценки тональных слов(2/2)

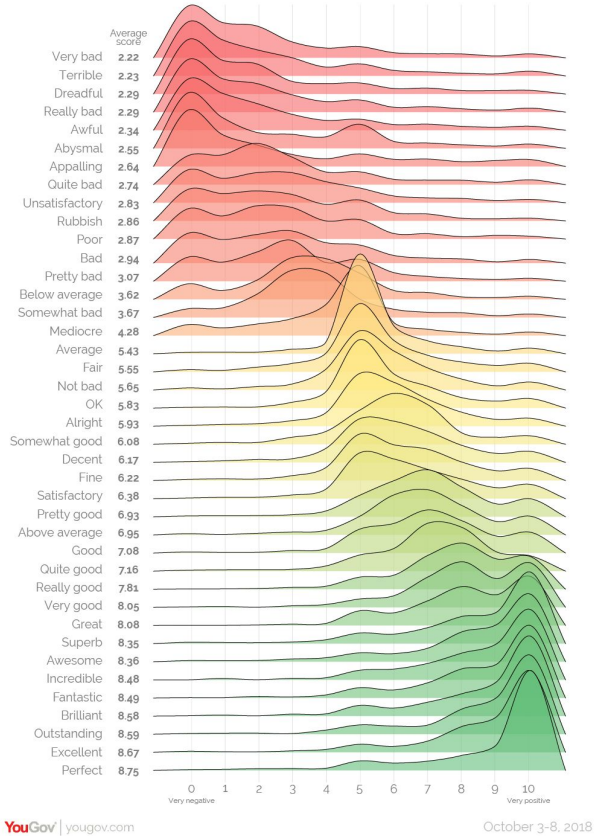


Двумерная шкала оценки

Границы изображены приблизительно

Симметрия предположительна

How “good” is good? (1/2)



Самое негативно окрашенное слово

“terrible” - 2.23 балла

Самое позитивно окрашенное слово

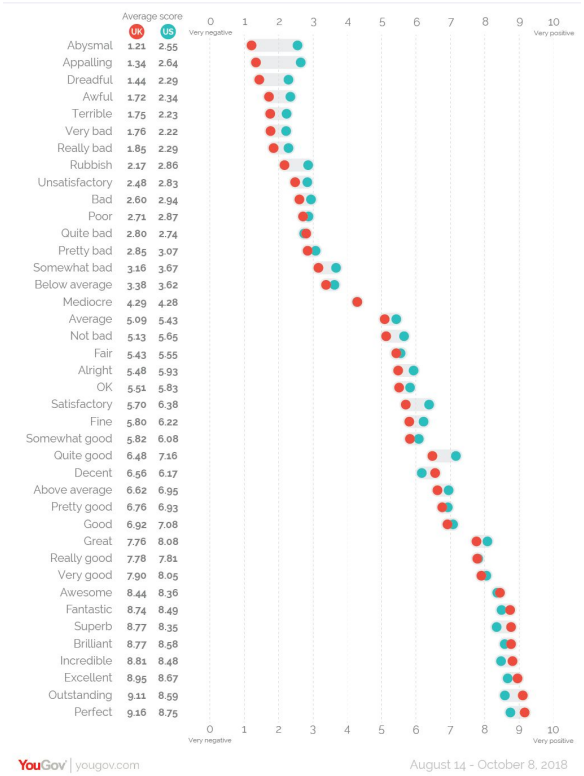
“excellent” - 8.85 баллов

Самое нейтрально окрашенное слово

“average” - 5.45 балла

2022 г.

How “good” is good? (2/2)



Самая существенная разница - US,UK

“abysmal” - 1.21 балл

“appalling” - 1.30 баллов

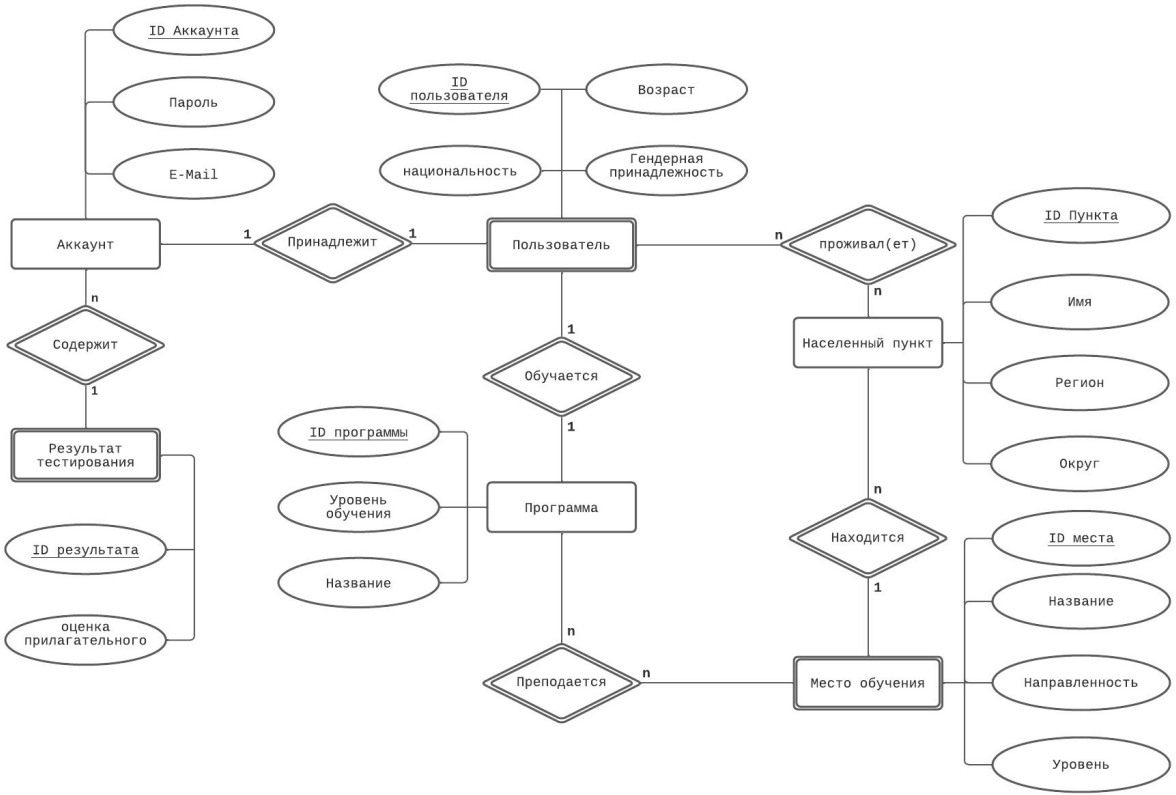
Самая небольшая разница - US,UK

“mediocre” - 0.01 балла

Будет ли результат аналогичного исследования по регионам России схожим?

2022 г.

Проектирование связей сущностей



Использование реляционных баз данных

Необходимо:
не только управление изолированными элементами,
но и учет связей между ними.

Характеристика реляционных БД	Почему не подходит к исследуемому набору данных
Сущности изолированы	Результат *Linked Object Mining* - набор *связанных* узлов
Предназначены для работы с таблицами	Набор данных *сильно взаимосвязан* и *слабоструктурирван*

Модели хранения noSQL (2/2)

Модель хранилища	Характеристика	Примеры
Колоночная	Данные хранятся в ячейках, сгруппированных в колонки. Колоночные семейства могут состоять из практически неограниченного количества колонок	Cassandra, HBase
Ключ-значение	Представляет собой большую хэш-таблицу	Redis, Voldemort
Графовая	Предназначена для хранения взаимосвязей и навигации в них	Amazon Neptune, Neo4j
Документо -ориентированная	Система «ключ-значение», которая позволяет к каждому ключу привязывать вложенные данные	MongoDB

Запросы к базе данных

Neo4j использует язык запросов Cypher.

```
MATCH (p:Product)-[:CATEGORY]->(l:ProductCategory)-[:PARENT*0..]->(:ProductCategory
{name:"Dairy Products"})
RETURN p.name
```

Тот же запрос на SQL:

```
SELECT p.ProductName
FROM Product AS p
JOIN ProductCategory pc ON (p.CategoryID = pc.CategoryID AND pc.CategoryName = "Dairy Products")

JOIN ProductCategory pc1 ON (p.CategoryID = pc1.CategoryID)
JOIN ProductCategory pc2 ON (pc1.ParentID = pc2.CategoryID AND pc2.CategoryName = "Dairy Products")

JOIN ProductCategory pc3 ON (p.CategoryID = pc3.CategoryID)
JOIN ProductCategory pc4 ON (pc3.ParentID = pc4.CategoryID)
JOIN ProductCategory pc5 ON (pc4.ParentID = pc5.CategoryID AND pc5.CategoryName = "Dairy Products");
```

Проектирование клиентского приложения

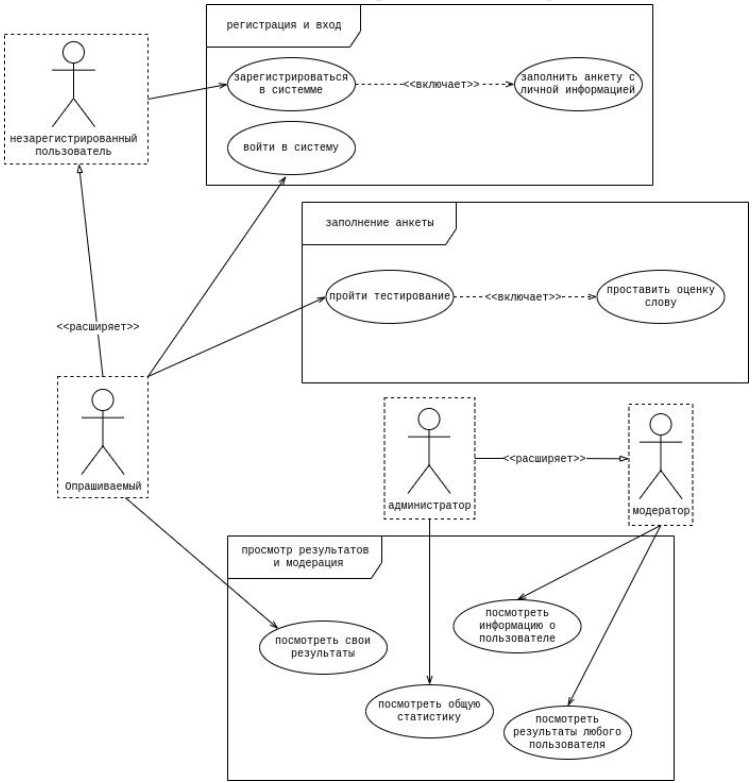


Диаграмма вариантов использования

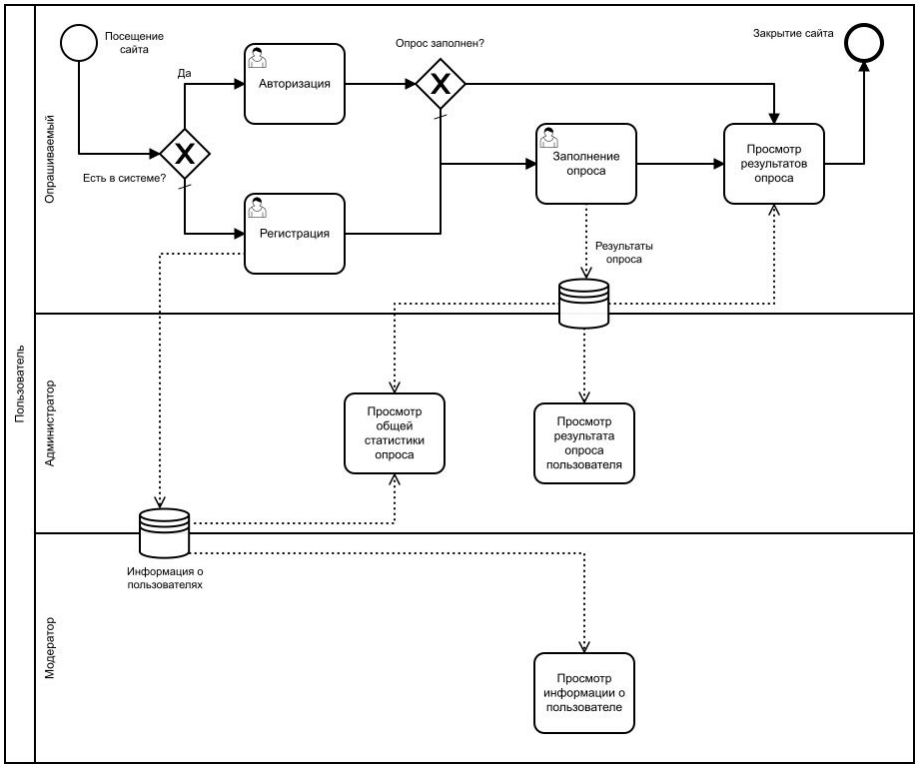


Диаграмма бизнес-процессов

Зависимость времени выполнения запросов от параметров системы

Время доступа к данным зависит от:

Размера кучи виртуальной машины Java и размера страничного кэша
(рассматривается в работе)

Соотношения объема кэшированных данных к общему объему хранимых данных

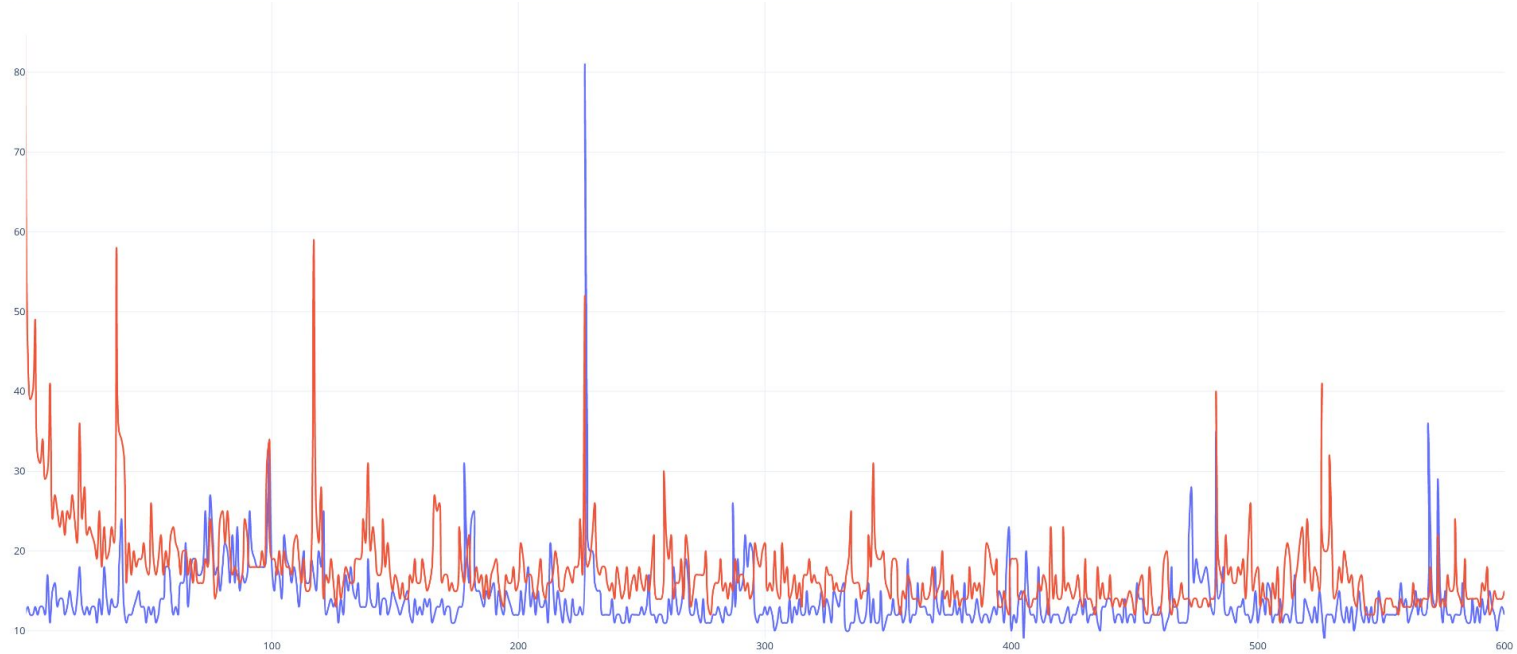
Скорости дискового накопителя: лучше использовать SSD или промышленный флеш-накопитель

Зависимость времени выполнения запросов от параметров системы

Увеличение размера кучи виртуальной машины Java	`dbms.memory.heap.initial_size` `dbms.memory.heap.max_size`
Увеличение размера страничного кэша	`dbms.memory.pagecache.size`

Зависимость времени выполнения запросов от параметров системы

Оранжевая линия – неоптимизированные параметры
Голубая линия – оптимизированные параметры



2022 г.

Заключение

Цель работы достигнута: спроектировано и реализовано программное обеспечение для проведения анкетированного опроса экспертов о тональности прилагательных.

выбрана наиболее подходящая модель для хранения данных, полученных в результате опроса - графовая;

спроектирована база данных, описаны ее сущности и связи;

разработана систему разметки интерфейса;

для выбранной графовой СУБД neo4j подобрана подходящая конфигурация, обеспечивающая повышение скорости доступа к данным.