

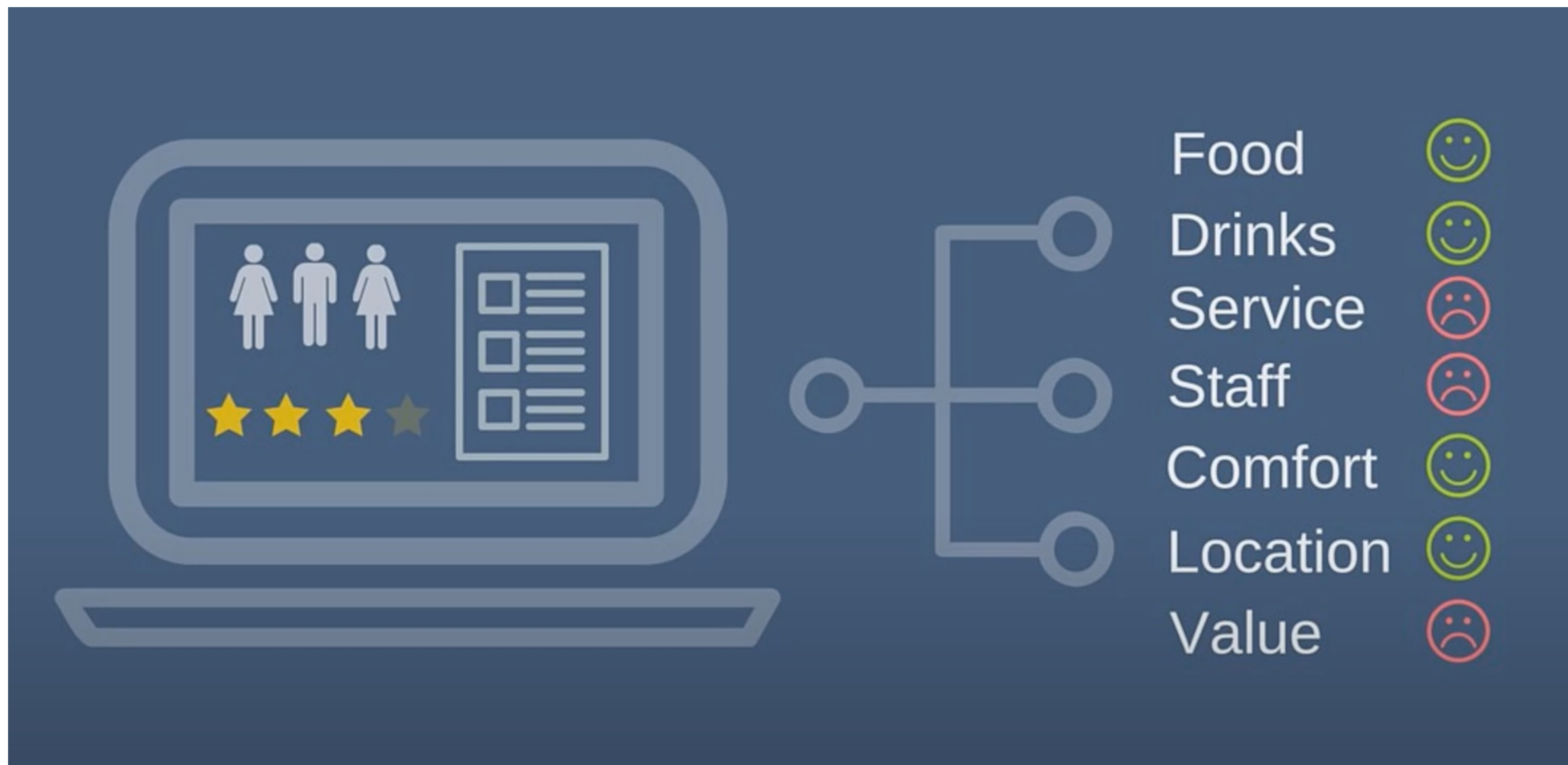
Аспектный анализ тональности

Докладчик: Аникевич Ю.В. 625 гр.

Научный руководитель: Ефремова Н.Э.

Аспектный анализ тональности

Сущность – объект мнения (продукт, сервис, тема, человек, организация, ...)







Аспектные категории

Аспектная категория (аспект) – характеристики (качества и свойства) заданной области

Например, для ресторана аспектами могут быть качество еды, обслуживания, интерьер. Для автомобиля - комфорт, надежность, внешний вид и прочее

5,0 Отлично
 1 726 отзывов

#1 из 15 в категории "отели" в регионе Кедеватан

	Расположение
	Чистота
	Обслуживание
	Цена/качество

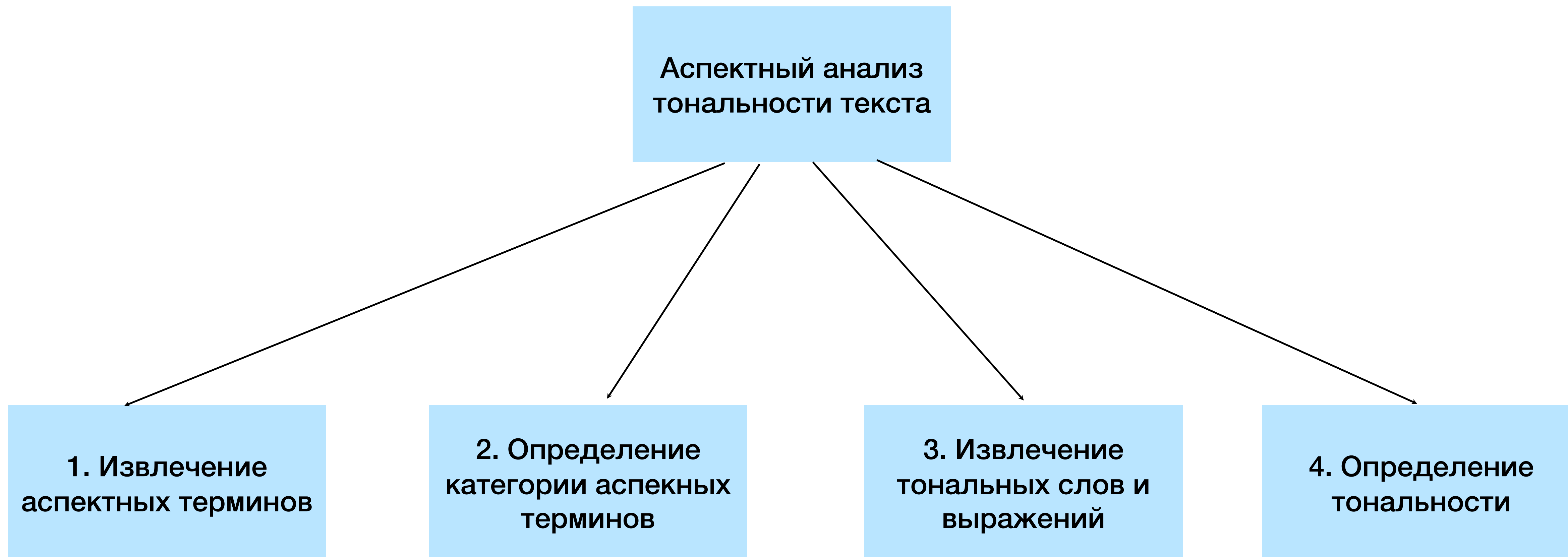
Аспектные термины

Для обозначения аспектов в текстах мнений используются аспектные термины

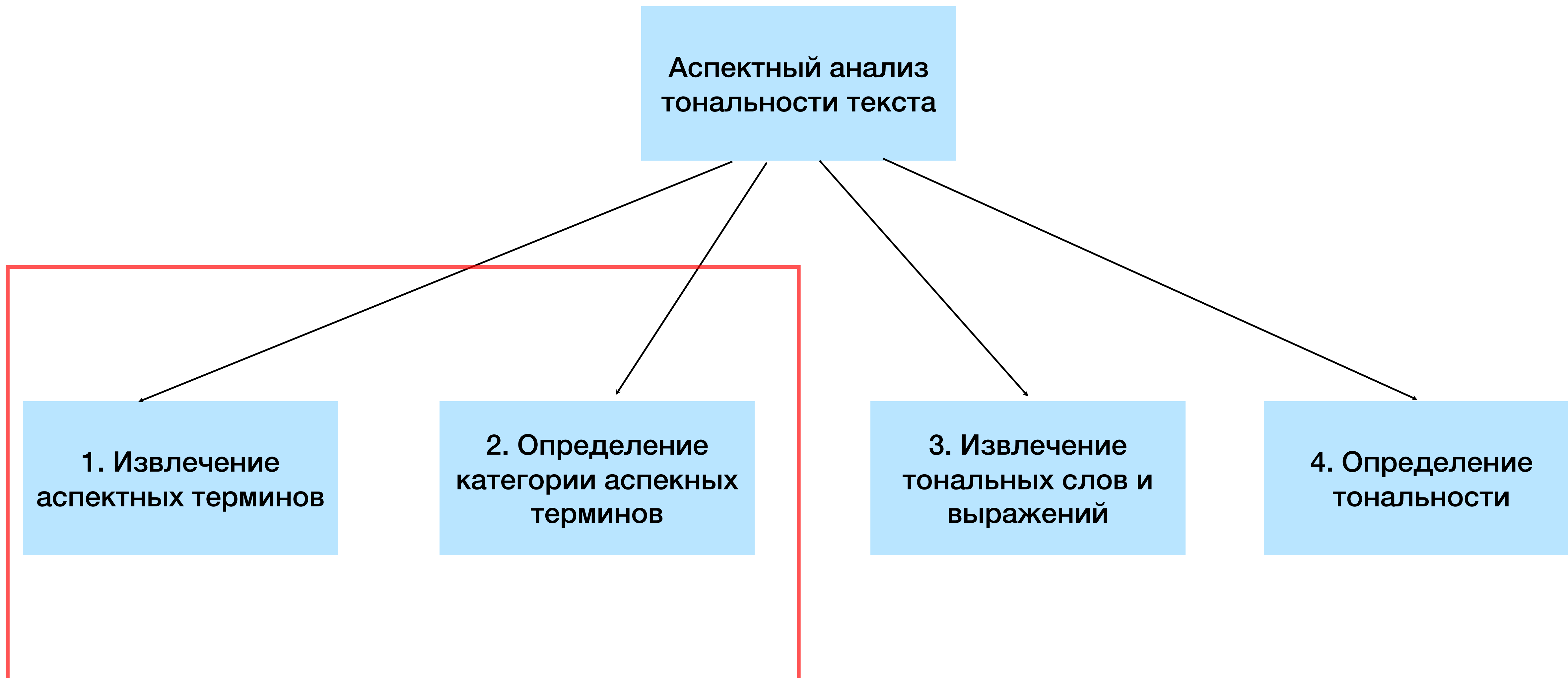
Аспектный термин - последовательность слов, относящихся к заданному аспекту объекта

- Еда очень вкусная
- Парковочные места все заняты
- Атмосфера этого ресторана средняя
- Обслуживание очень медленное, но официантка была вежлива
- Пицца стоит очень дорого

Подзадачи



Подзадачи



Извлечение аспектных терминов

Можно выделить три подхода для извлечения аспектных терминов:

1. Лингвистический подход

- основан на лингвистических правилах и шаблонах

2. Статистический подход

- извлечения слов и словосочетаний, наиболее часто встречающихся в документах

3. Машинное обучение

- бинарная классификация слов на аспектные и неаспектные
- разметка последовательности (например, BIO-кодирование)

Статистический подход. Основные шаги

Метод статистического подхода, как правило, содержит следующие шаги:

1. Выявление существительных и именных групп.
2. Вычисление для них статистической характеристики.
 - *наиболее часто используемые*
 - *TF-IDF*
 - *C-value*
3. Отбор именных словосочетаний, значение статистической характеристики у которых выше заданного порога.

Статистический подход. Метод N&L

Алгоритм:

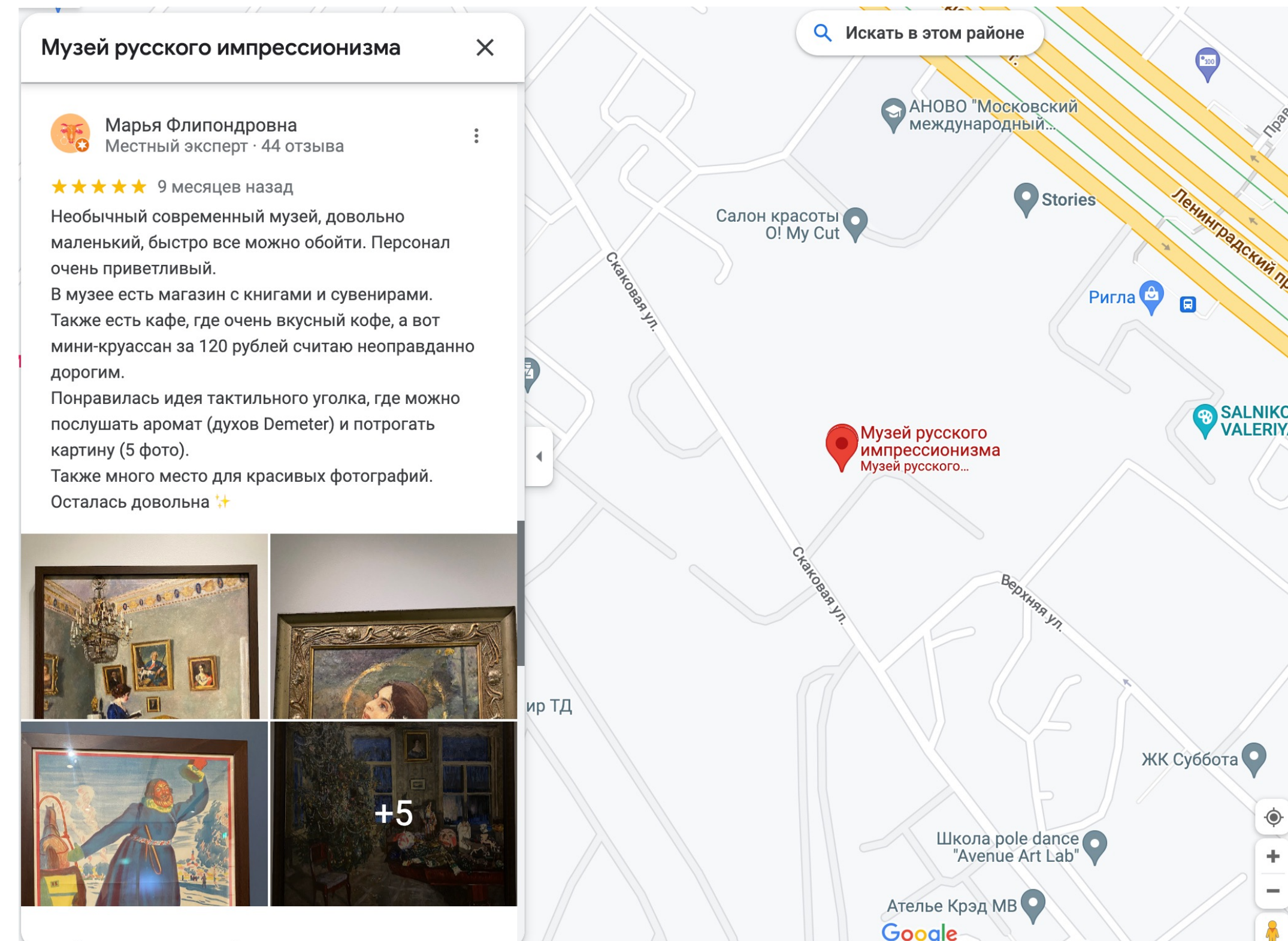
1. Выделение всех существительных и именных групп
2. Добавление новых кандидатов в аспектные термины. Формируются пары и триплеты из кандидатов, выявленных на шаге 1 в рамках одного предложения в том же порядке

Пример: *срок службы + батарея = срок службы батареи*

3. Вычисляется значение ***p-support*** для каждого кандидата.
p-support для термина t – это количество предложений, содержащих t , исключая предложения, содержащие другой потенциальный отдельный аспектный термин t' , который включает t .
4. Отсечение всех кандидатов, ниже некоторого заданного значения p -support.

Набор данных

- 180 музеев
 - *Музей Охоты И Рыболовства*
 - *Музей иллюзий*
 - *Политехнический музей*
 - *Музей-квартира А. С. Пушкина*
 - *и т.д.*
- 174к отзывов
- Использован Google Maps Reviews Crawler (<https://www.botsol.com/bots/google-maps-reviews-crawler>)



Часто используемые термины

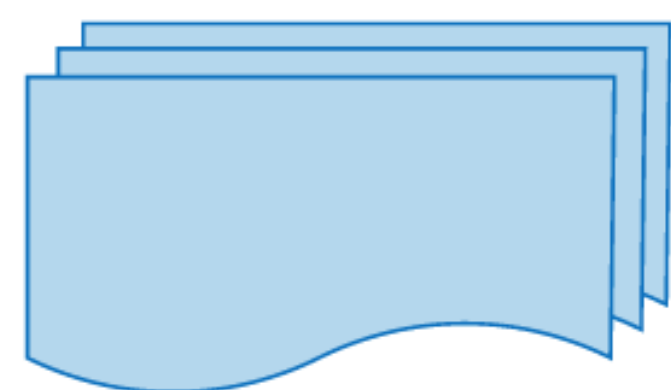
Термин	Количество использований
музей	31706
ребёнок	18586
парк	18304
экскурсия	16675
билет	15306
дворец	13274
посещение	13217

Термин	Количество использований
вход	12289
фонтан	12065
история	11717
человек	9148
очередь	8930
экскурсовод	7804
цена	7430

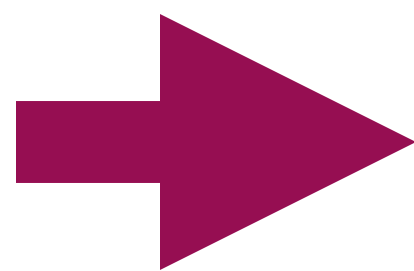
Выделение аспектных категорий

- **Цель** - определение аспектных категорий заданной области (не известны заранее)
- **Решение:**
 - Выделение аспектных терминов
 - Кластеризация через семантическое сходство (Word2Vec)
- **Результат:** N аспектных категорий с соответствующим списком аспектных терминов

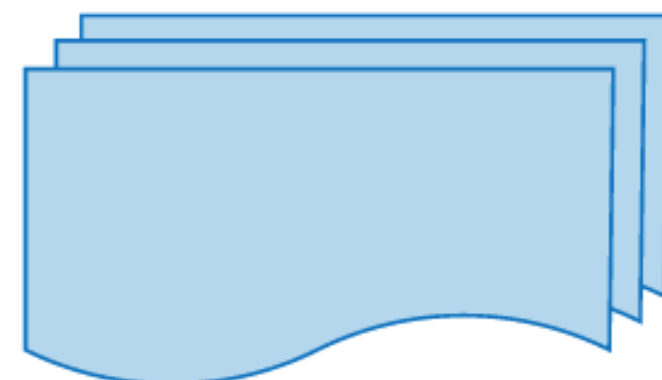
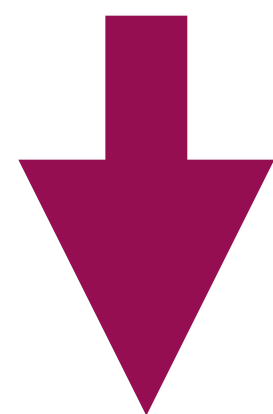
Последовательность шагов



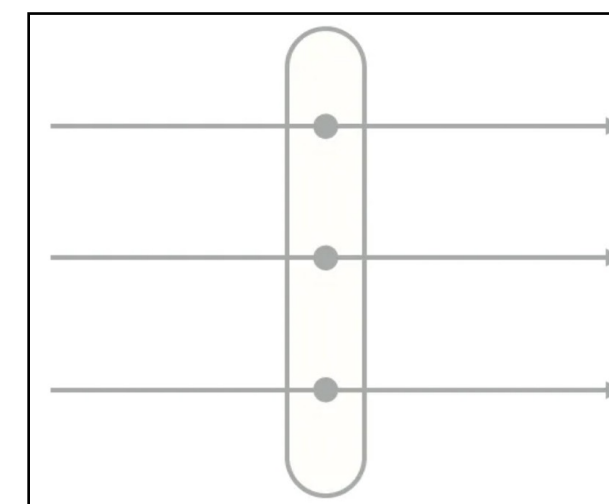
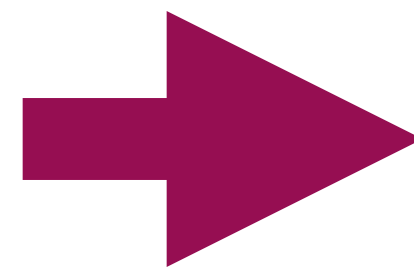
Корпус отзывов



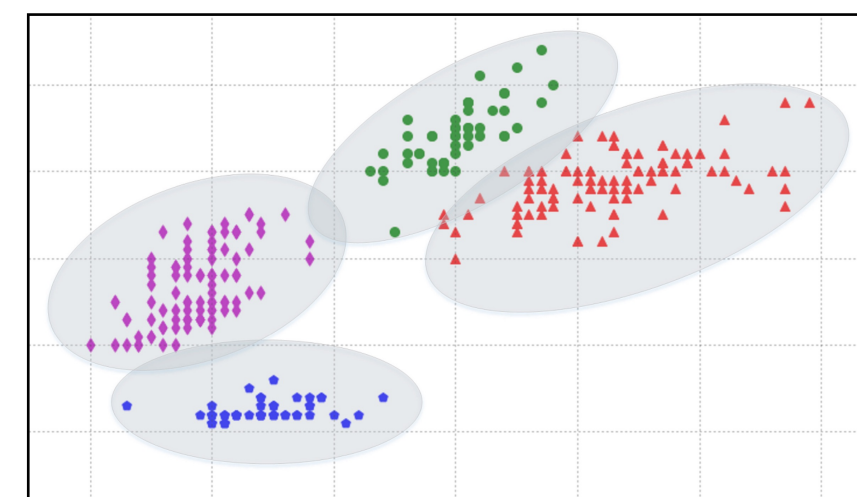
Модуль
извлечения
аспектных
терминов



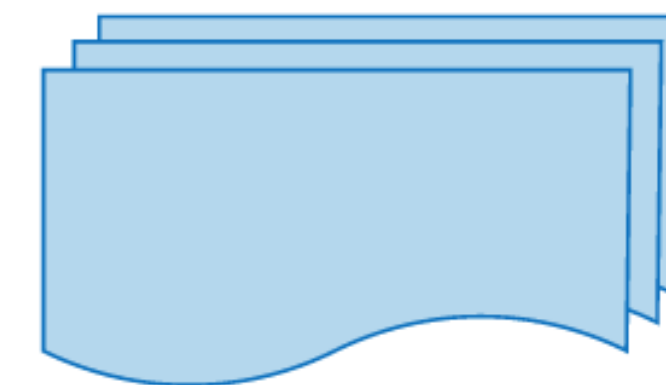
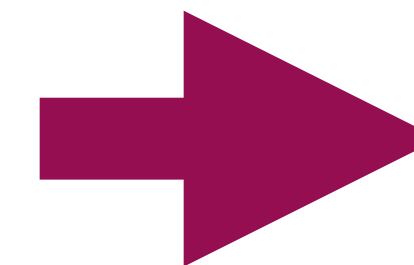
Список аспектных
терминов



Word2Vec



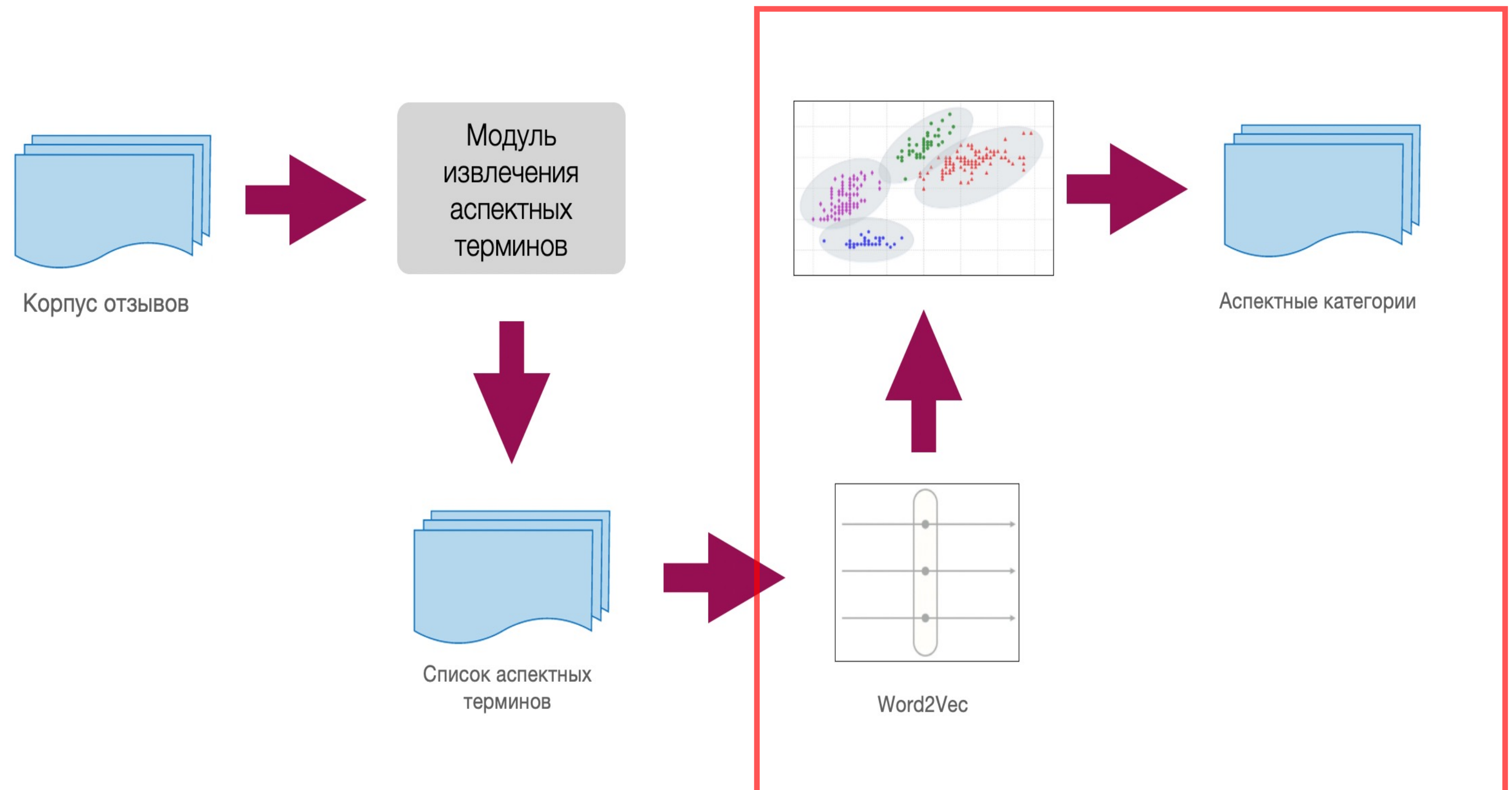
Кластеризация



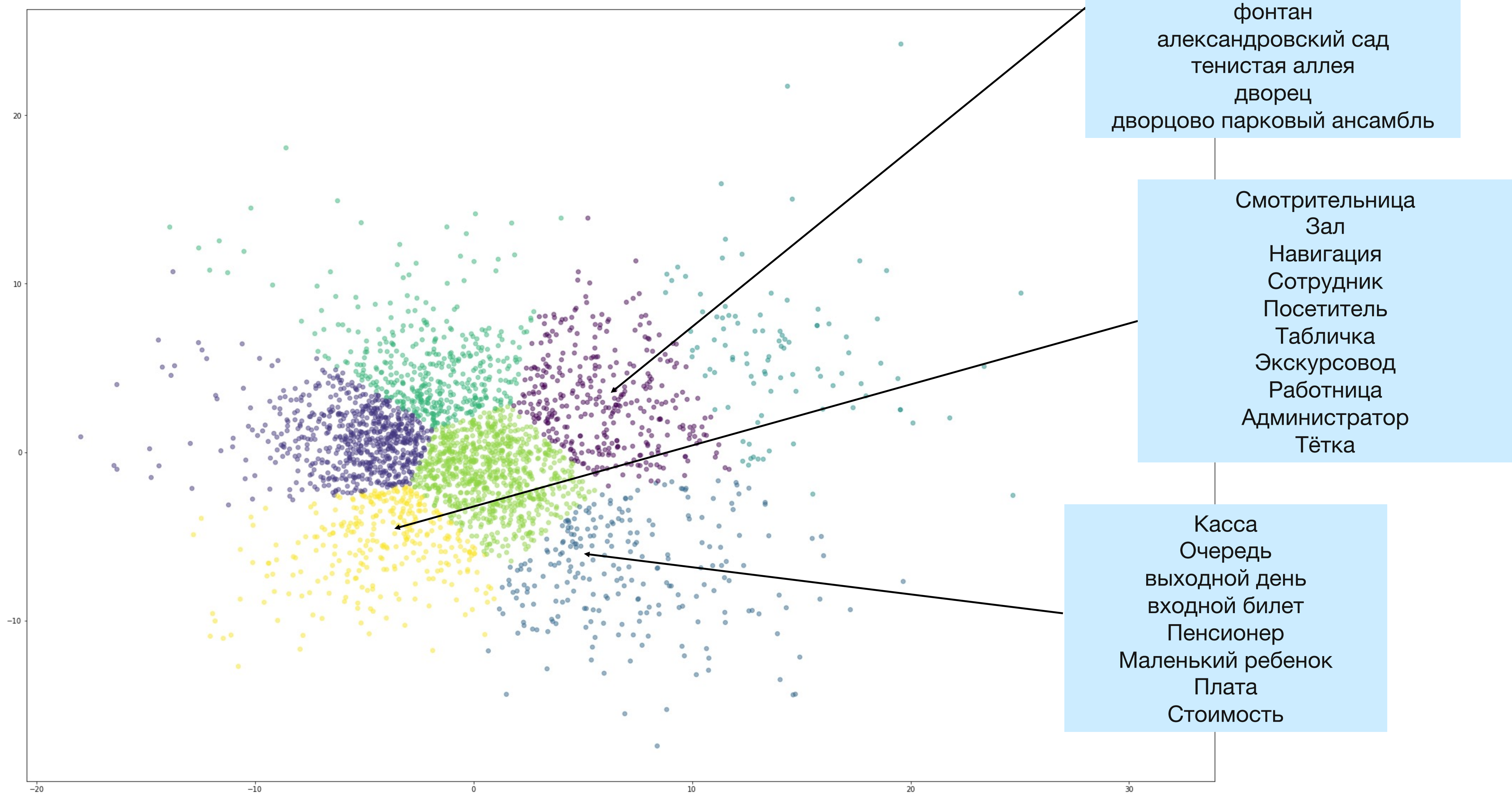
Аспектные категории

Алгоритм кластеризации

- Векторное представление Word2Vec
- Модуль кластеризации
 - K-means
 - Метод локтя
- На выходе список терминов для каждой аспектной категории
- Не зависит от рассматриваемой сущности



Результаты



Дальнейшая работа

1. Улучшение реализованных алгоритмов
2. Рассмотрение следующих подзадач:
 - Извлечение тональных слов и выражений
 - Определение тональности