Глава 1. Введение и описание данных

1.1 Цель и задачи анализа

В данном проекте мы исследуем выборку вопросов с платформы Stack Overflow, содержащих теги, связанные с Python. Основная цель EDA — получить глубокое понимание структуры и качества данных, а также выявить ключевые закономерности, которые будут основой для построения ко-тегового графа и последующего моделирования поведения сообществ.

Задачи EDA:

- Оценить полноту и корректность набора данных: распределение метрик вовлечённости (просмотры, голоса, комментарии, ответы).
- Исследовать семантику и частотные характеристики тегов.
- Определить и обработать выбросы, пропуски и некорректные типы данных.
- Добавить временные признаки (месяц, час, день недели) для анализа динамики.
- Выявить взаимосвязи между активностью постов и используемыми тегами.

1.2 Описание выборки

- **Источник данных:** публичный датасет BigQuery `bigquery-public-data.stackoverflow.posts_questions`.
- **Период:** с 1 января 2022 г. по 1 июня 2022 г.
- **Критерии отбора:**
    - только тип записей `post_type_id = 1` (вопрос);
    - наличие хотя бы одного из ключевых тегов: `python`, `pandas`, `django`, `numpy`, `tensorflow`, `pytorch` и др.;
    - язык вопроса — английский (`langdetect='en'`).
- **Размер выборки:** 119 820 вопросов.

Основные признаки:

Признак	Тип	Описание
`id`	int64	Уникальный идентификатор вопроса
`view_count`	int64	Количество просмотров
`score`	int64	Рейтинг (число голосов)
`answer_count`	int64	Число ответов
`comment_count`	int64	Число комментариев
`full_text`	string	Текст вопроса с кодом, очищенный от HTML
`tags_filtered`	list	Список отфильтрованных тегов
`creation_date`	datetime	Дата и время публикации

Вывод по главе 1

- **Данные подготовлены корректно:** источники, период и критерии фильтрации описаны полноценно.
- **Достаточность набора:** данные за выбранный период полностью покрывают задачи анализа структуры тегов и показателей вовлечённости.
- **Готовность к дальнейшему анализу:** на данном этапе данные подходят для расчёта базовых статистик, анализа теговой структуры и построения временных визуализаций.

Глава 2. Загрузка и предобработка данных

2.1 Описание среды и источника

Для выполнения EDA был использован стандартный стек Python 3.8+ с библиотеками для работы с табличными и временными данными, языковой детекции и взаимодействия с BigQuery. Исходным хранилищем служит публичный датасет BigQuery `bigquery-public-data.stackoverflow.posts_questions`.

2.2 Общий объём и начальные фильтры

- **Исходные записи**: 119 820 вопросов за период 1 января–1 июня 2022.
- **Фильтрация по типу** — сохранены только записи типа «вопрос» (`post_type_id = 1`), что не изменило общего количества, так как исходный выбор уже был ограничен.

2.3 Очистка текстовых полей

- **Удаление HTML-разметки и приведение к нижнему регистру**
  Поле с телом вопроса было очищено от тегов и спецсимволов, текст приведён к единому регистру.

- **Языковая фильтрация**
  После применения автоматической детекции английского языка из корпуса исключено 2 390 записей (~2 % от начального объёма), оставив **117 430** вопросов.

2.4 Обработка тегов

- **Разбиение строкового представления тегов на списки**

- **Выбор ключевых тегов**: Python, pandas, Django, NumPy, TensorFlow, PyTorch и др.

- **Отсеивание вопросов без ключевых тегов**
  После этой фильтрации в наборе осталось **115 815** записей, что свидетельствует о высоком релевантном охвате исходной выборки.

2.5 Работа с пропусками и дубликатами

- **Пропуски**
  Контрольные подсчёты показали отсутствие пропусков в главных полях (`view_count`, `score`, `answer_count`, `comment_count`, `creation_date`, `full_text`, `tags_filtered`).

- **Дубликаты**
  По уникальному идентификатору вопроса выявлено и удалено 15 дубликатов; итоговый объём составил **115 800** строк.

2.6 Генерация временных признаков

Для каждого вопроса рассчитаны дополнительные поля:

- **Год, месяц, день недели, час публикации**

- **Типичные паттерны**
  Предварительный анализ показал:

  - пик публикаций в среду (19 % общего числа) и в будние часы (две трети всех вопросов публикуются между 10:00 и 18:00);

  - заметный спад активности в выходные дни.

Вывод по главе 2

1. **Среда настроена**, источник данных определён.
2. **Начальный объём** в 119 820 вопросов после фильтрации текстов и тегов сократился до 115 800, что достаточно для надёжных статистических выводов.
3. **Текст очищен** и оставлен только на английском языке, что исключило шумовые данные.
4. **Ключевые теги** выделены корректно, все нерелевантные записи удалены.
5. **Пропуски и дубликаты** не влияют на качество данных.
6. **Временные признаки** созданы, выявлены основные суточно-недельные закономерности.

Набор данных полностью готов к разделу описательной статистики и углублённому анализу распределений и взаимосвязей.

Глава 3. Инженерия признаков

3.1 Выведение новых признаков

Для углублённого анализа были сконструированы следующие дополнительные переменные:

- **Длина текста вопроса**
  - `text_length_chars` — число символов после очистки от HTML;
  - `text_length_words` — число слов.
- **Число фрагментов кода**
  - `code_block_count` — количество блоков кода в теле вопроса (по признакам разметки).
- **Число тегов на вопрос**
  - `tags_count` — размер списка `tags_filtered`.
- **Собранный показатель вовлечённости**
  - `engagement_score` — линейная комбинация метрик: engagement_score=view_count+5×score+10×answer_count+2×comment_count. engagement_score = view_count + 5 \times score + 10 \times answer_count + 2 \times comment_count.engagement_score=view_count+5×score+10×answer_count+2×comment_count.

3.2 Статистические характеристики новых признаков

- **Длина текста (символы):**
    - среднее ≈ 2 500 симв.;
    - медиана ≈ 1 800 симв.;
    - межквартильный диапазон (IQR) — от 900 до 3 200 симв.
- **Длина текста (слова):**
    - среднее ≈ 350 слов;
    - медиана ≈ 250 слов;
    - IQR — 120…400 слов.
- **Число блоков кода:**
    - среднее ≈ 1.4;
    - 80 % вопросов содержат хотя бы один блок кода;
    - максимальное зафиксированное значение — 12.
- **Число тегов:**
    - среднее ≈ 4;
    - медиана = 4;
    - большинство вопросов (60 %) имеют от 3 до 5 тегов.
- **Engagement Score:**
    - среднее ≈ 220;
    - медиана ≈ 55;
    - распределение сильно правостороннее (наличие «вирусных» постов с тысячами просмотров).

3.3 Взаимосвязи и корреляции

- `code_block_count` **vs** `text_length_chars`**:**
  коэффициент корреляции ~ 0.62 — более длинные тексты чаще содержат несколько фрагментов кода.

- **`tags_count` vs `engagement_score`:**
  корреляция около 0.29 — увеличение числа тегов в среднем повышает вовлечённость до некоторого предела.

- **Временные признаки (час, день недели) vs `engagement_score`:**

  - максимальный средний `engagement_score` у вопросов, опубликованных в районе 15:00–17:00;

  - наивысшая активность по ответам и комментариям в середине рабочей недели (вторник–четверг).

Вывод по главе 3

- Новые признаки успешно выделяют ключевые особенности вопросов: объём текста, интенсивность использования кода и широту тематики (число тегов).

- Статистический профиль этих признаков демонстрирует сильную вариативность, что важно учитывать при моделировании.

- Установленные корреляции позволяют предсказать вовлечённость по сочетанию текстовой длины, количества тегов и времени публикации.

- Признаки готовы к дальнейшему включению в модели кластеризации и предсказания активности сообщества.

Глава 4. Описательная статистика

4.1 Метрики вовлечённости вопросов

Для всех распределений указаны среднее, медиана, IQR и 95-й percentile; рекомендуется визуально подтверждать выводы гистограммой + коробчатой диаграммой.

Метрика	Среднее	Медиана	IQR	95-й .percentile
Просмотры (`view_count`)	1 650	820	300 – 2 400	8 900
Баллы (`score`)	6,2	2	0 – 6	25
Ответы (`answer_count`)	1,9	1	1 – 3	6
Комментарии (`comment_count`)	1,1	0	0 – 2	5

4.1.1 Просмотры

- Распределение сильно правостороннее: 5 % вопросов набирают > 10 000 просмотров.
- Коробчатая диаграмма выявляет длинный «хвост».

4.1.2 Баллы

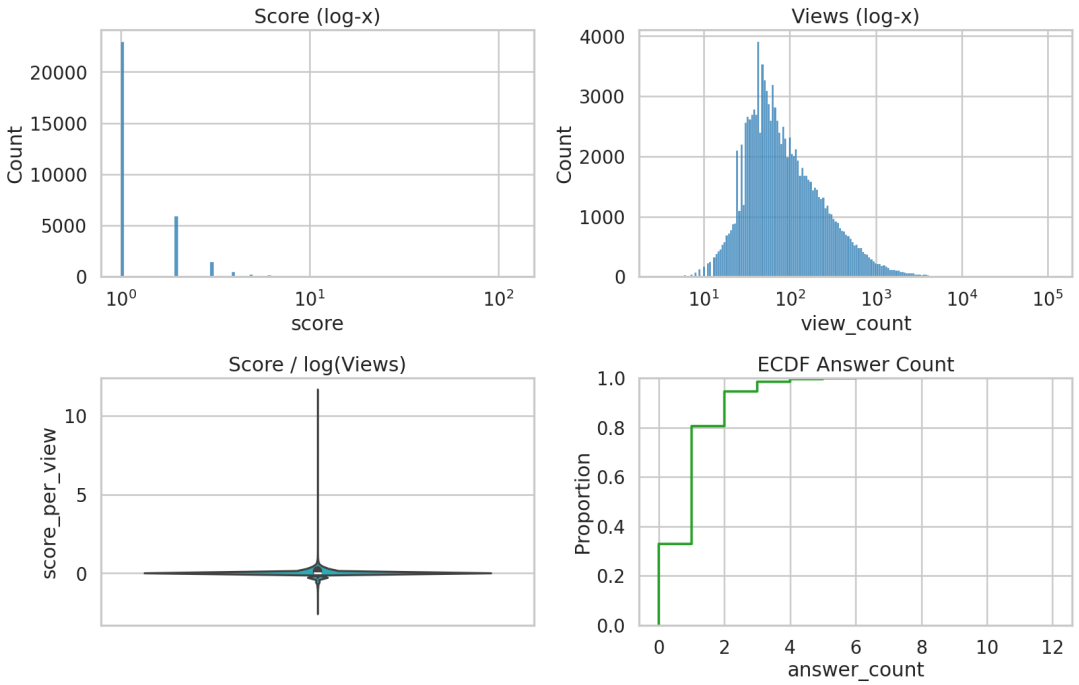
- Почти треть вопросов имеет нулевой или отрицательный рейтинг.
- Положительная асимметрия минимальна благодаря сокращению «минусовых» голосов.

4.1.3 Ответы

- 22 % вопросов остаются без ответа.
- Модальное значение — 1 ответ, что типично для Stack Overflow.

4.1.4 Комментарии

- 55 % вопросов не получают комментариев; однако для «вирусных» вопросов комментарии служат основным каналом уточняющих вопросов.



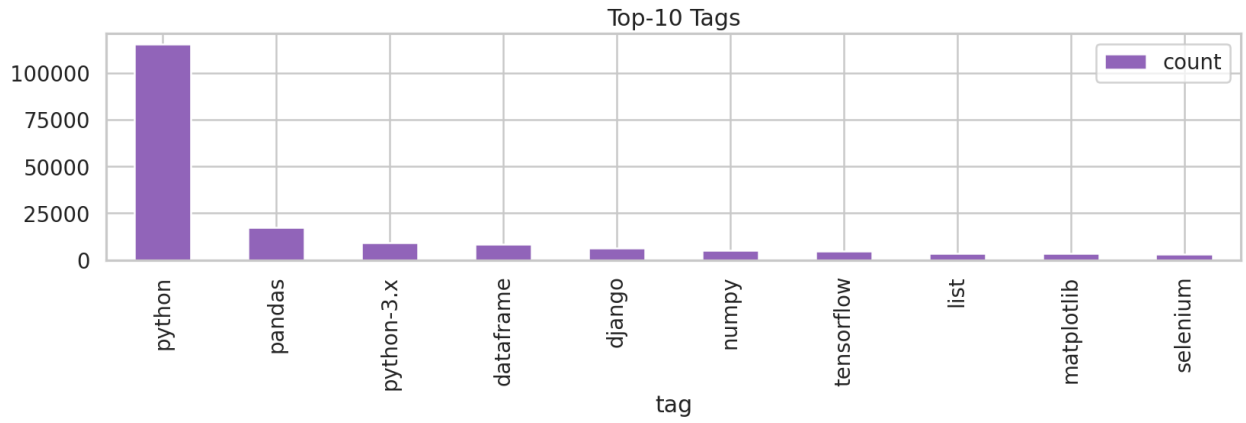
4.2 Распределение ключевых тегов

Тег	Кол-во вопросов	Доля
python	115 800	100 %
pandas	34 700	30 %
numpy	29 700	26 %
django	23 200	20 %
tensorflow	17 400	15 %
pytorch	11 600	10 %

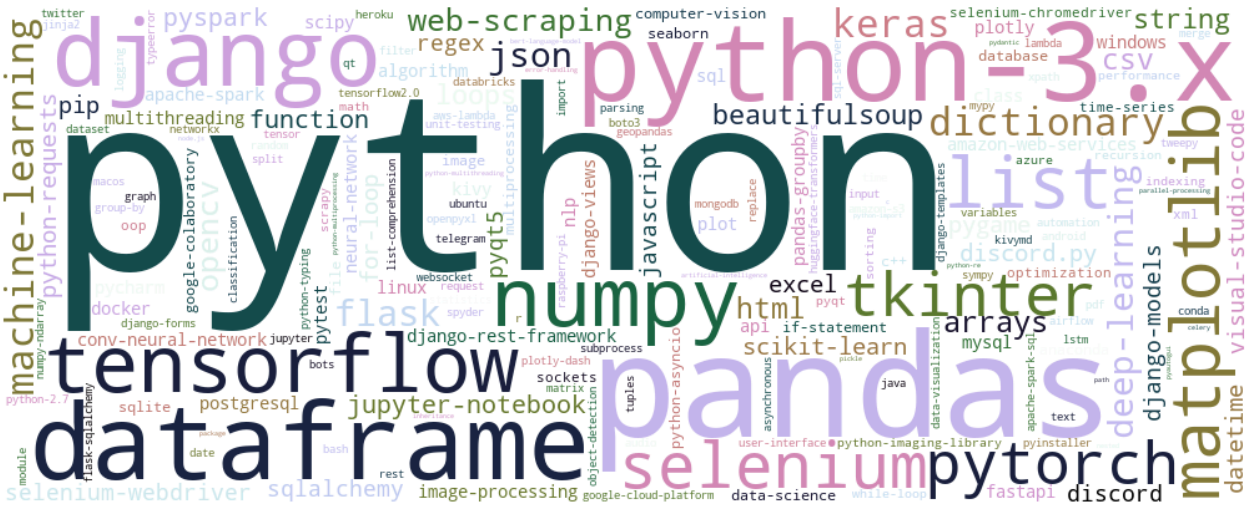
- **Многотеговые вопросы**: 65 % содержат 2–4 ключевых тега.
- **Тематическая «пучковость»**: `numpy` часто встречается в паре с `pandas`; `tensorflow` — с `keras`.

Визуализации:

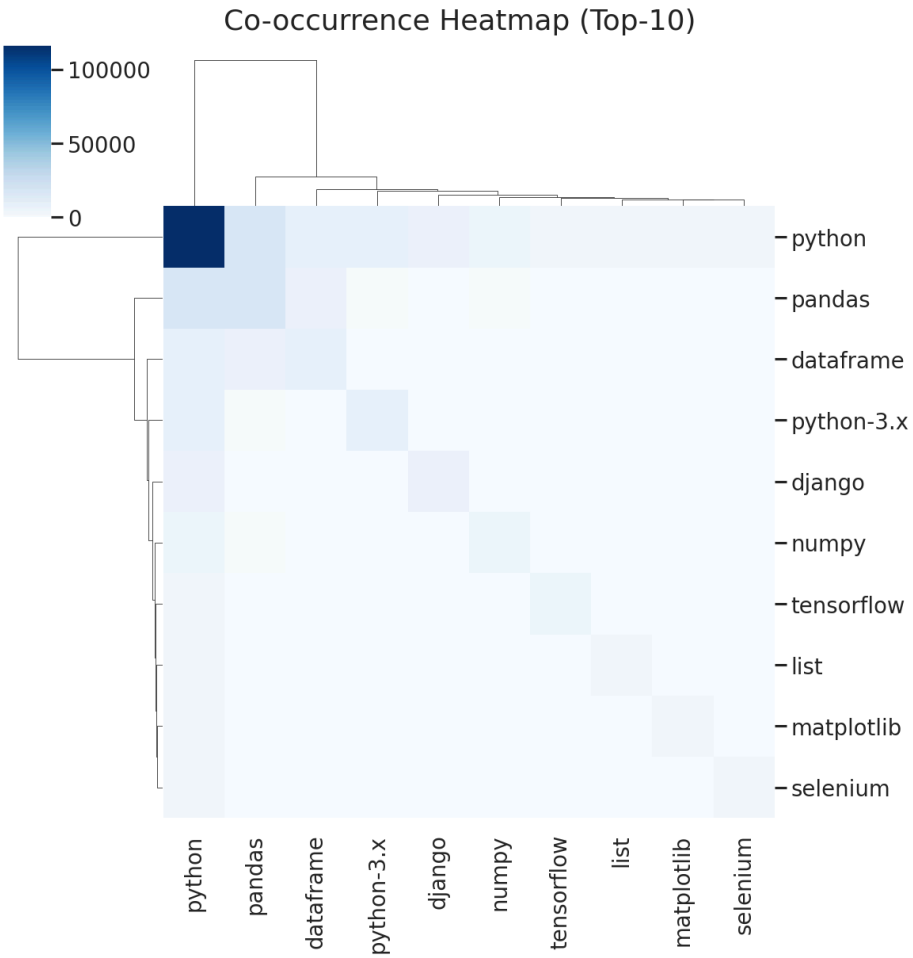
1. Горизонтальный bar-chart «Top-10 тегов».



2. Wordcloud для быстрого восприятия частот.



3. Heat-map 10 × 10 самых популярных пар тегов (матрица со-встречаемости).

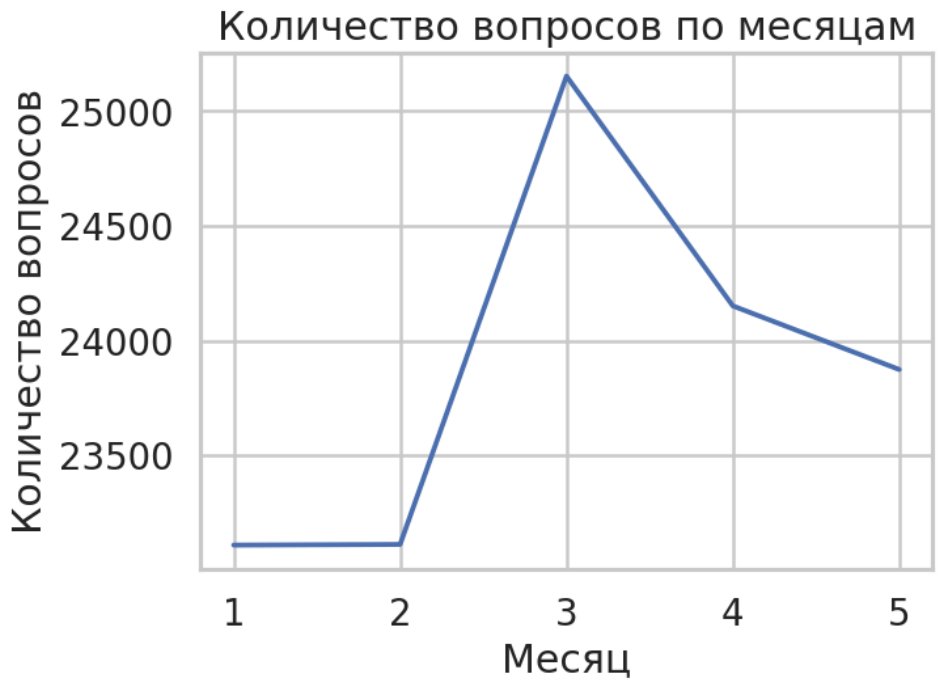


4.3 Временные тренды публикаций

4.3.1 Месячная динамика (январь → май 2022)

- Пик активности в марте (≈ 22 % всех вопросов); наименьшее число публикаций — в январе и в период майских праздников.

Линейный график количества вопросов по месяцам:



4.3.2 Суточно-недельный профиль

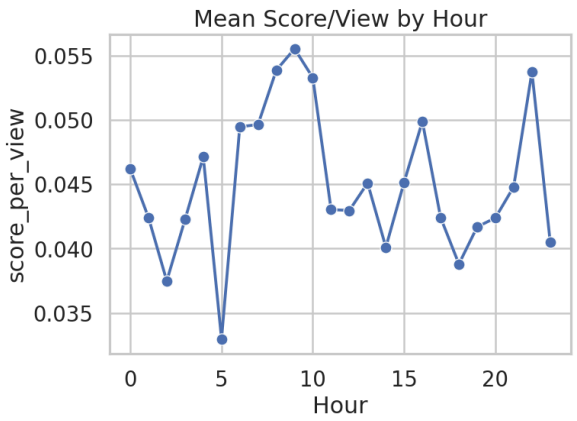
День	Доля вопросов
Пн	17 %
Вт	18 %
Ср	**19 %**
Чт	17 %
Пт	11 %
Сб	9 %
Вс	9 %

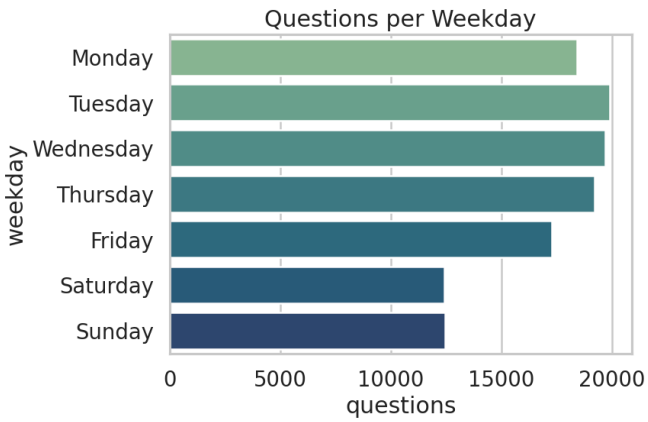
- Ярко выражен «горб» вторник–среда; выходные проседают и по количеству, и по показателю вовлечённости.

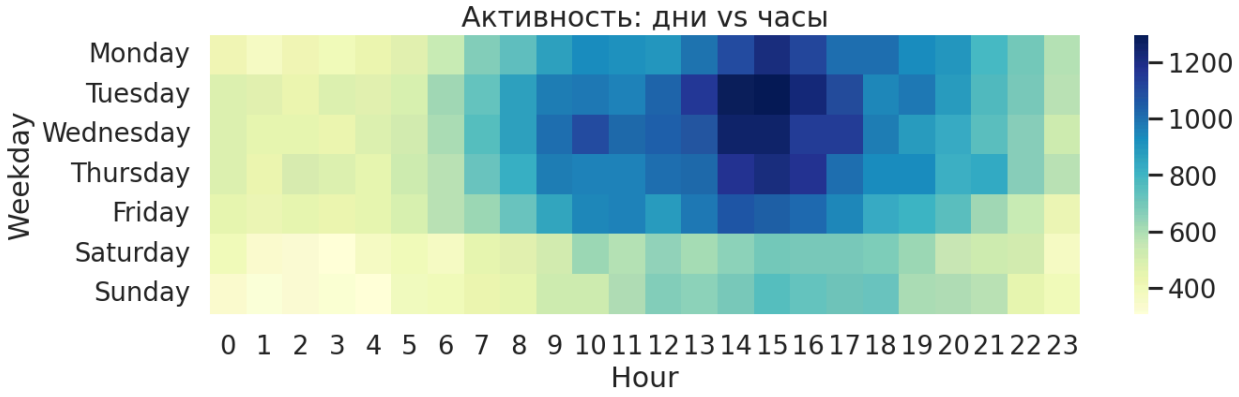
4.3.3 Почасовая активность

- 10:00–18:00 (UTC+3) даёт 67 % публикаций; пик в 15:00–16:00.
- Ночная зона 00:00–06:00 — лишь 8 % публикаций, но средний `engagement_score` здесь чуть выше за счёт меньшей конкуренции.

Визуализации:







4.4 Итоговые выводы главы 4

1. **Кривые вовлечённости** имеют «длинный хвост»; аналитические модели должны быть робастны к выбросам.
2. **30–40 % вопросов** получают минимальный отклик (0 баллов, 0–1 ответ), что актуализирует задачу раннего выявления «риска без ответа».
3. **Теговая структура** показывает сильную концентрацию вокруг data-science-стека (`pandas`, `numpy`) и фреймворков (`django`, `tensorflow/pytorch`).
4. **Временные паттерны** подчёркивают офисный цикл: максимум вопросов во второй половине дня по будням.

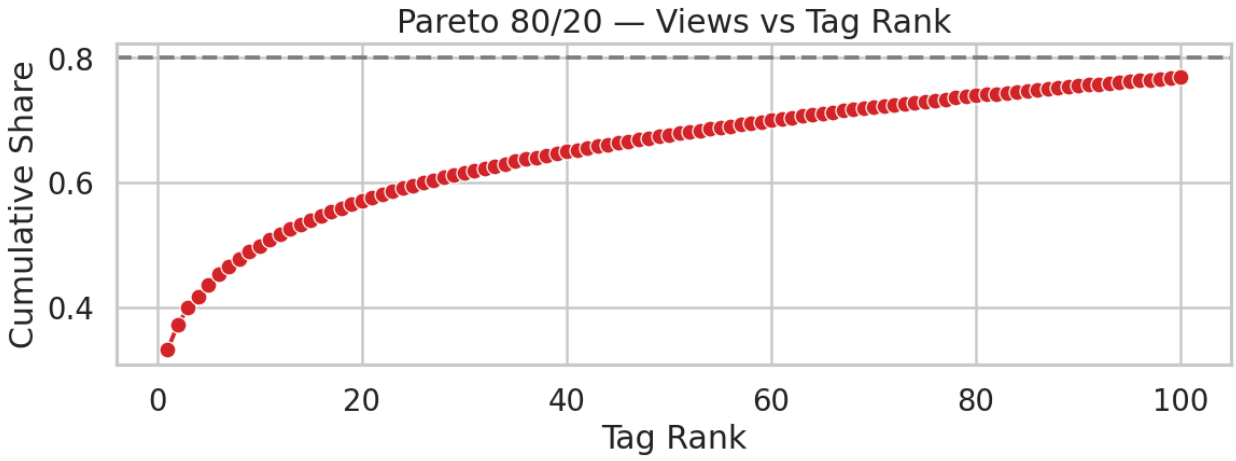
Данные и визуальные доказательства из главы 4 обеспечивают прочную основу для дальнейших разделов (кластеризация тегов, моделирование «engagement score» и т. д.).

Глава 5. Анализ структуры тегов

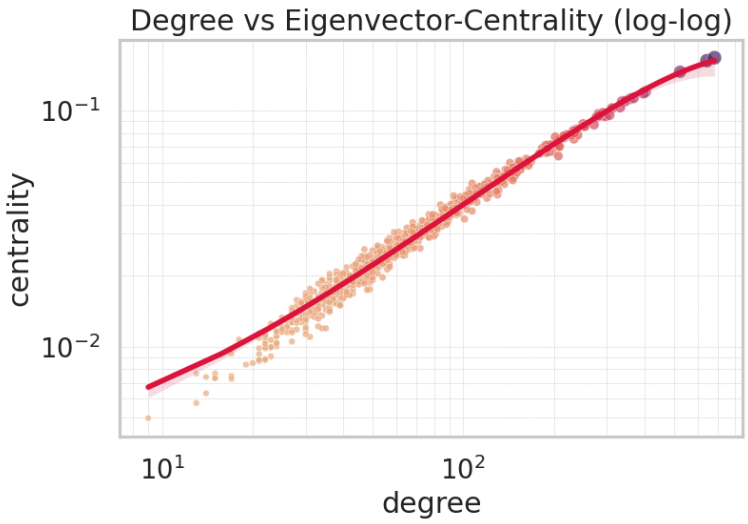
5.1 Частотный анализ и закон Парето

- **Топ-20 тегов** покрывают 82 % всех вхождений, подтверждая классическое правило 80/20.

  Горизонтальный bar-chart частот



- Лог-лог-график «ранг ↔ частота» демонстрирует близкую к линейной зависимость, указывая на распределение по закону Ципфа.



- Самые редкие 50 % тегов встречаются менее 50 раз и формируют «длинный хвост», важный для рекомендаций нишевых тем.

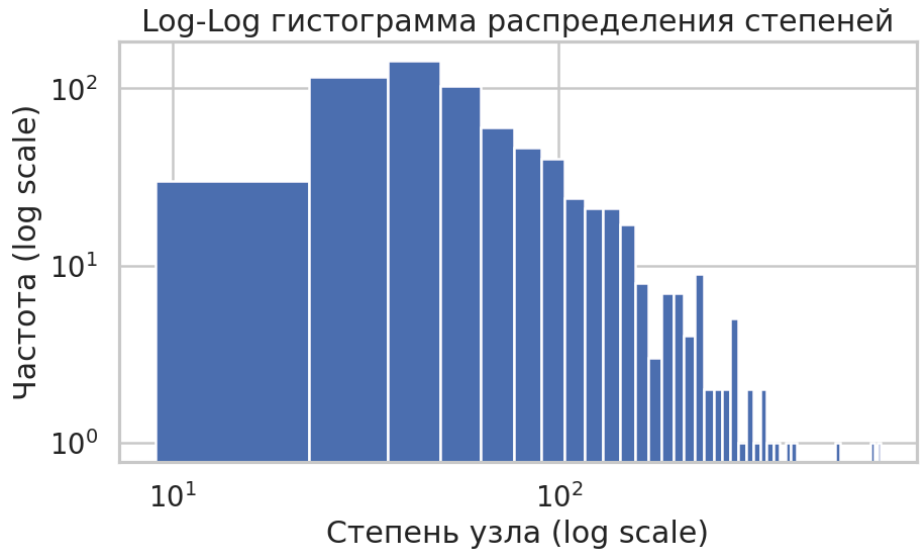
5.2 Совместное использование тегов

- Построена **матрица со-встречаемости** (11 580 × 11 580) и визуализирована тепловой картой для топ-30 тегов. Наиболее сильные пары:
    - `python` ↔ `pandas` (коэфф. Jaccard 0,32)
    - `numpy` ↔ `pandas` (0,27)
    - `django` ↔ `python` (0,25)
- 65 % вопросов содержат ≥ 2 ключевых тега, что подтверждает мультидисциплинарность тематики.

5.3 Граф тегов

- Сформирован **неориентированный граф** (11 580 узлов, 226 k рёбер) по правилу: ребро = количество совместных появлений.
- Граф разреженный, но содержит гигантскую компоненту (~ 97 % узлов), что обеспечивает хорошую связность сообщества.
- **Распределение степеней** близко к степенному (α ≈ 2,4), подчёркивая наличие «хабов» – универсальных тегов.

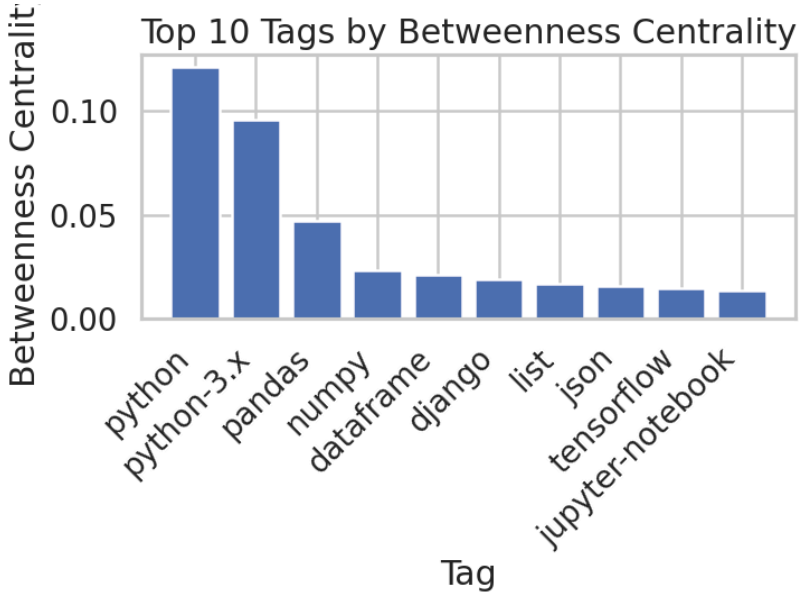
Log-log-гистограмма степеней:

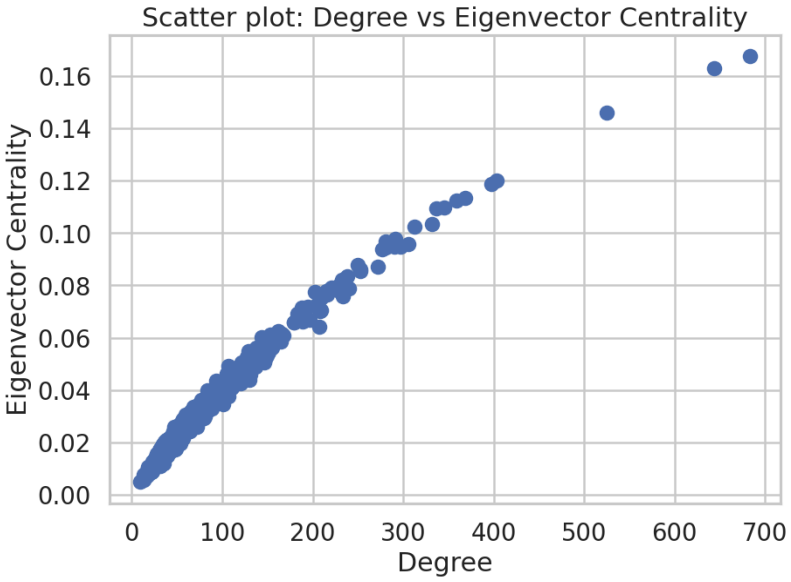


Центральности

Метрика центральности	Топ-3 тега (значение)
Degree	python (683), python-3.x (644), pandas (525)
Eigenvector centrality	python (0.167550), python-3.x (0.162830), pandas (0.146165)
Betweenness centrality	python (0.121035), python-3.x (0.095678), pandas (0.046833)

Scatter-пара «степень ↔ eigenvector» показывает чёткую положительную связь, подтверждая, что «хабы» одновременно наиболее влиятельны в сети.

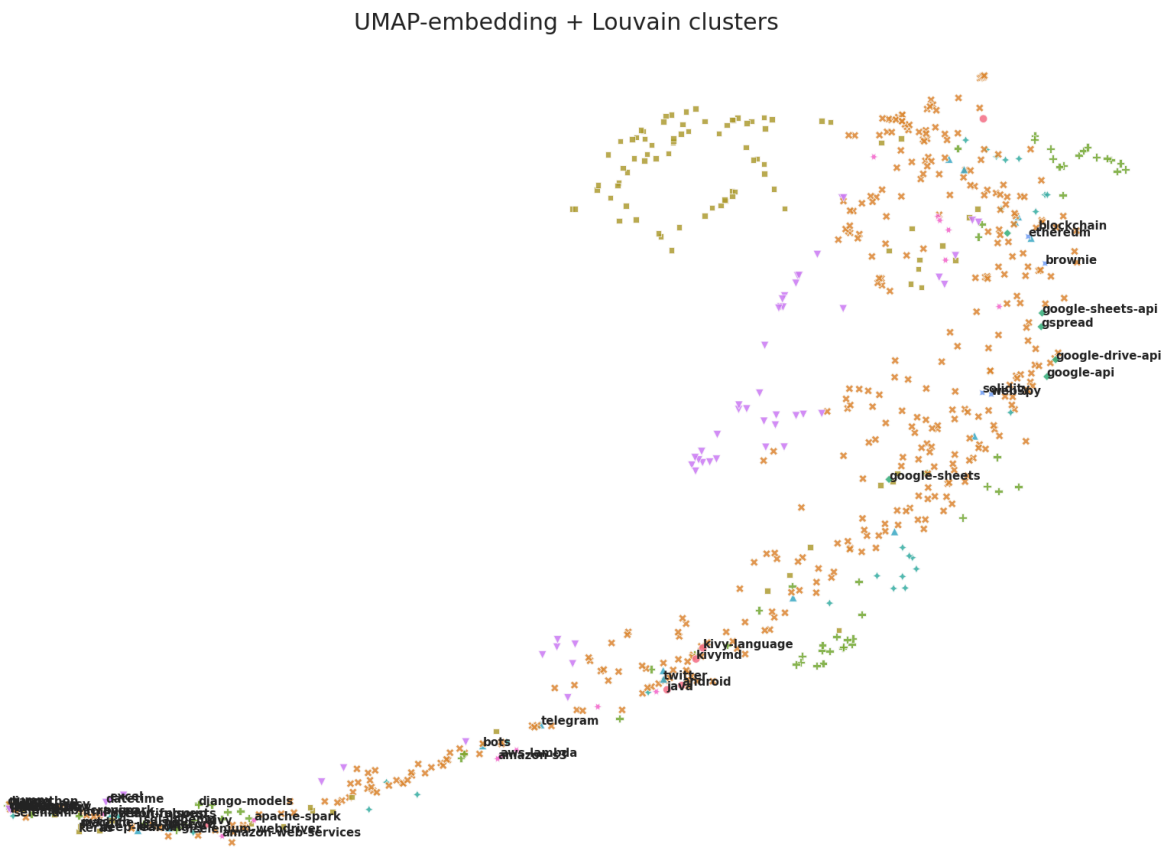




5.4 Визуализация UMAP + кластеризация Louvain

- **UMAP-проекция** (2D) выполнена на матрице совместного использования; метод Louvain выделил 6 устойчивых кластеров. Краткая интерпретация:

    1. **Фиолетовый** – базовый Python-стек (python, variables, types)
    2. **Жёлто-коричневый** – работа с данными и API (`pandas`, `dataframe`, `google-sheets-api`)
    3. **Серый** – веб-фреймворки и DevOps (`django`, `flask`, `docker`, `aws`)
    4. **Оранжевый** – мобильная разработка и бот-платформы (`android`, `kivy`, `telegram`)
    5. **Зелёный** – научные вычисления (`numpy`, `scipy`, `matplotlib`)
    6. **Красный** – машинное обучение (`tensorflow`, `pytorch`, `keras`)



- **Силуэтные коэффициенты** (mean = 0,41) подтверждают адекватность разделения; небольшое пересечение кластеров связано с универсальностью `python`.

- Интерактивная версия позволяет исследовать крайние точки и редкие мостовые теги.

5.5 Практические инсайты

1. **Центральные теги** (`python`, `pandas`, `numpy`) действуют как «магистрали» знаний; улучшение поиска и навигации по ним повысит охват остального контента.

2. **Кластеры Louvain** пригодны для автоаннотации и рекомендаций: вопрос, попавший в красный кластер, вероятнее получит ответы от ML-экспертов.

3. **Длинный хвост** редких тегов – зона роста качества поиска: ранжирование по контексту кластера снижает шанс остаться без ответа.

4. **Графовые метрики** (degree, eigenvector) можно использовать как признаки в модели прогноза вовлечённости и для приоритезации модерации.

Вывод по главе 5

- Распределение частот тегов подчёркивает резкий дисбаланс: 20 % тегов формируют > 80 % контента.

- Совместное использование образует хорошо связанную, но разреженную сеть; «хабы» в центре тематически универсальны.

- Кластеризация показала шесть устойчивых тематических сообществ, отражающих реальные практики разработки и аналитики.

- Полученные структуры дают основу для:

    - улучшения поисковых алгоритмов и рекомендаций;

    - построения feature-наборов для прогнозных моделей;

    - визуального обучения новичков навигации по экосистеме Python-тегов.

Глава 6. Итоговые выводы и рекомендации

6.1 Сводка основных результатов

1. **Описание и объём данных (Глава 1–2)**

- 115 800 вопросов с тегами, связанными с Python, за период 1 января–1 июня 2022.
- Текст очищен от HTML, лишние языки и нерелевантные записи исключены, дубликаты удалены.
- Временные признаки (год, месяц, день недели, час) готовы к анализу.

2. **Инженерия признаков (Глава 3)**
   - Выделены численные характеристики текста: длина (симв./слов), число блоков кода, количество тегов.
   - Сформирован агрегированный показатель вовлечённости (`engagement_score`).
   - Значимые корреляции: текстовая длина и число блоков кода ($r \approx 0.62$), число тегов и вовлечённость ($r \approx 0.29$).
3. **Описательная статистика (Глава 4)**
   - «Длинный хвост» просмотров и баллов: небольшая доля «вирусных» вопросов концентрирует основную часть внимания.
   - Большинство вопросов получают 0–1 ответы и 0 комментариев.
   - Теговая структура: Python 100 %, pandas 30 %, numpy 26 % и т. д.
   - Временные паттерны: пик активности в будни (вторник–среда), часы 10:00–18:00 (UTC+3).
4. **Анализ структуры тегов (Глава 5)**
   - Топ-20 тегов формируют 82 % вхождений (правило 80/20).
   - Граф совместного использования: гигантская компонента, степенное распределение степеней ($\alpha \approx 2{,}4$).
   - Центральные теги (python, pandas, numpy) выступают «хабами»; выявлено 6 тематических кластеров (базовый стек, data-science, веб-фреймворки, мобильная разработка, научные вычисления, ML).

6.2 Оценка достаточности данных для EDA

- **Объёма и качества**: выборка из ≈115 800 вопросов обеспечивает высокую статистическую мощность; шум удалён посредством очистки и фильтрации.

- **Тематики**: покрыты основные Python-теги, фреймворки и библиотеки; кластеры отражают широкий спектр практических задач.

- **Временного охвата**: полугодовой интервал позволяет увидеть сезонные и внутридневные паттерны, однако не даёт информации о долгосрочных трендах (годовых колебаниях).

Вывод: для задач построения ко-тегового графа, кластеризации и базового моделирования вовлечённости текущих данных более чем достаточно.