Задача 1. *Аналитик в компании "Рога и Ко" за день написал некоторое количество SQL-запросов. Все запросы можно охарактеризовать количественной характеристикой сложности выполнения – условная "память". Три "тяжелейших" (7/20 общей затраченной памяти) запроса превысили максимальное время выполнения и были отключены системой. Три "легчайших" (5/13 затраченной памяти всех оставшихся запросов) выполнялись менее 10 минут каждый. Остальные запросы выполнялись менее получаса каждый.*

1. *Найдите общее количество запросов, которое написал аналитик*
2. *Дайте оценку на среднее время успешно выполненного запроса*

Ответ. Всего аналитик написал 10 запросов. Для среднего времени $\overline{m}$ можно дать следующую оценку: $\frac{40}{7} < \overline{m} < \frac{150}{7}$.

Решение. *Пункт 1. Общее число запросов*
Введём следующие обозначения:
n – общее число запросов;
m_i – сложность выполнения i-го запроса (его "память");
M – общая затраченная память.
Пусть нумерация запросов произведена так, что $m_1 \leqslant m_2 \leqslant \ldots \leqslant m_n$. Тогда из условия задачи получим:

$$\begin{cases} m_1 + m_2 + \ldots + m_n = M \\ m_{n-2} + m_{n-1} + m_n = \frac{7}{20}M \\ m_1 + m_2 + m_3 = \frac{5}{13}(M - \frac{7}{20}M) = \frac{5}{20}M \end{cases}$$

Значит,

$$\frac{5}{20}M + m_4 + \ldots + m_{n-3} + \frac{7}{20}M = M$$

$$m_4 + \ldots + m_{n-3} = \frac{8}{20}M$$

Назовём запросы $m_4, \ldots, m_{n-3}$ *промежуточными*. Известно, что если 3 запроса занимают $\frac{7}{20}$ памяти, они будут отключены системой. Значит, промежуточных запросов точно не меньше четырёх (иначе получится, что 3 запроса занимают $\frac{8}{20}$ всей памяти.
С другой стороны, знаем, что именно запросы m_1, m_2, m_3, занимающие $\frac{5}{20}$ памяти, названы легчайшими. Значит, никакие 3 процесса из промежуточных в совокупности не занимают меньше $\frac{5}{20}M$, потому что иначе были бы включены в группу легчайших. Покажем, что если промежуточных запросов будет 5, это условие не сможет быть выполнено.
Интуитивное доказательство: Если 5 запросов занимают $\frac{8}{20}M$, то в среднем один запрос занимает $\frac{8/5}{20}M$, а значит три запроса в сумме занимают $\frac{24/5}{20}M < \frac{5}{20}M$. Значит, не получится при таком количестве запросов сделать так, чтобы никакие три из них не занимали меньше $\frac{5}{20}M$.
Строгое доказательство: Переименуем сложности промежуточных запросов в a, b, c, d, e. Тогда $a+b+c+d+e = \frac{8}{20}M$. Рассмотрим все возможные комбинации троек: таких всего $C_5^3 = \frac{5!}{3!2!} = 10$. Хотим, чтобы сумма элементов внутри каждой тройки превосходила $\frac{5}{20}$:

$$a + b + c > \frac{5}{20}M$$

$$a + b + d > \frac{5}{20}M$$

$$a + b + e > \frac{5}{20}M$$

$$\ldots$$

$$c + d + e > \frac{5}{20}M$$

Сложим все суммы, получим:

$$6(a + b + c + d + e) > 10 * \frac{5}{20}M$$

$$6(a + b + c + d + e) > \frac{50}{20}M$$

Поскольку $a + b + c + d + e = \frac{8}{20}M$, то $6(a + b + c + d + e) = \frac{48}{20} \not> \frac{50}{20}$. Доказали. Осталось проверить, что если промежуточных запроса 4, все условия могут быть выполнены. Пусть все промежуточные запросы весят одинаково, тогда вес каждого составляет $\frac{8/4}{20}M = \frac{2}{20}$. Значит, сумма любых трёх составит $\frac{6}{20} > \frac{5}{20}$.
Таким образом, промежуточных запросов было 4, а общее количество запросов $n = 3 + 4 + 3 = 10$.

Пункт 2. Среднее время выполненного запроса
Поскольку всего было создано 10 запросов, а 3 были отключены системой, то успешно выполнились всего 7 запросов. Про них нам известно:

$$0 < m_1, m_2, m_3 < 10$$
$$10 < m_4, m_5, m_6, m_7 < 30$$

Значит, верно следующее:

$$40 < \sum_{i=1}^{7} m_i < 150$$
$$40 < \overline{m} * 7 < 150$$
$$\frac{40}{7} < \overline{m} < \frac{150}{7}$$

Задача 2. *В команде есть два стажера-аналитика. Правильный ответ каждый из стажеров получает в 14 из 17 случаев. Чтобы быть более уверенным в важном решении, менеджер решил дать одну и ту же задачу сразу обоим аналитикам: если оба получают одинаковый ответ, то менеджер его использует, а если ответы разные – выберет один из них наугад. Насколько такой способ повышает шансы менеджера принять верное решение?*

Ответ. Шансы принять верное решение останутся прежними.

Решение. Рассмотрим все возможные случаи и посчитаем вероятности:

1) Оба аналитика правы: $\frac{14}{17} * \frac{14}{17} = \frac{196}{289}$
2) Первый прав, а второй нет: $\frac{14}{17} * \frac{3}{17} = \frac{42}{289}$
3) Второй прав, а первый нет: $\frac{3}{17} * \frac{14}{17} = \frac{42}{289}$
4) Оба дали неверный ответ: $\frac{3}{17} * \frac{3}{17} = \frac{9}{289}$

Посчитаем вероятность, с которой менеджер выберет верное решение:

$$\frac{196}{289}*\mathbf{1} + \frac{42}{289}*\frac{\mathbf{1}}{\mathbf{2}} + \frac{42}{289}*\frac{\mathbf{1}}{\mathbf{2}} + \frac{9}{289}*\mathbf{0} = \frac{196}{289} + \frac{42}{289} = \frac{238}{289} = \frac{14}{17}$$

При слагаемых $\frac{42}{289}$ стоит коэффициент $\frac{1}{2}$, поскольку при разных ответах менеджер выбирает наугад, а значит с вероятностью $\frac{1}{2}$ выберет верный ответ.

Задача 3. *В мешке лежат три кубика: 6-гранный, 12-гранный, 20-гранный. Мы достали один кубик наудачу, подкинули его и на нем выпало 4. Какова вероятность, что если мы так же достанем и подкинем один из оставшихся в мешке кубиков, на нем выпадет меньше?*

Ответ. 30%.

Решение. Рассмотрим все возможные случаи и посчитаем вероятности:

1) Пусть вытащили 6-гранный кубик. Далее, с вероятностью $\frac{1}{2}$ мы вытащим 12-гранный и с вероятностью $\frac{1}{2}$ – 20-гранный. Вероятность того, что на 12-гранном кубике выпадет число, меньшее 4, равна $\frac{3}{12}$, а на 20-гранном – $\frac{3}{20}$. Таким образом, вероятность того, что мы вытащим кубик, значение на котором меньше 4, равна:
$$p_1 = \frac{1}{2} * \frac{3}{12} + \frac{1}{2} * \frac{3}{20} = \frac{1}{5}.$$
2) Пусть первым вытащили 12-гранный кубик:
$$p_2 = \frac{1}{2} * \frac{3}{6} + \frac{1}{2} * \frac{3}{20} = \frac{13}{40}.$$
3) Пусть первым вытащили 20-гранный кубик:
$$p_3 = \frac{1}{2} * \frac{3}{6} + \frac{1}{2} * \frac{3}{12} = \frac{3}{8}.$$

Поскольку все вышеперечисленные ситуации равновероятны, то искомая вероятность равна:

$$P = \frac{1}{3} * p_1 + \frac{1}{3} * p_2 + \frac{1}{3} * p_3 = \frac{1}{3} * \frac{1}{5} + \frac{1}{3} * \frac{13}{40} + \frac{1}{3} * \frac{3}{8} = \frac{13}{40} = \frac{3}{10}$$

Задача 4. *Чтобы между пользователями Авито было больше доверия, а жизнь мошенников стала сложнее, мы решили попробовать ввести систему отзывов: покупатель может оставить отзыв на продавца. Отзыв может быть просто рейтингом (1-5 звездочек), а может содержать дополнительно какой-то произвольный текст.*

1. *Предложите метрики, по которым можно будет следить за прогрессом такого проекта и определять его успешность.*
2. *Поскольку Авито не магазин, а площадка для связи покупателя и продавца, мы в большинстве случаев не знаем, произошла ли в действительности сделка и на каких условиях. Для большинства сделок последнее, что нам известно – покупатель нажал кнопку просмотра телефона продавца или связался с ним в чате. Также мы всегда знаем логин (привязанный к email и телефону) продавца, но покупатель может быть незалогиненным. В связи с этим кажется, что есть большой риск накрутки отзывов и недобросовестного использования системы: например, профессиональные участники будут пытаться оставлять негативные отзывы на своих конкурентов и хвалебные на себя. Предположим, что система некоторое время уже работает и у нас есть данные по отзывам и всей активности клиентов: что продавал, что искал, на каких объявлениях смотрел телефоны и т.п.*
   Как оценить масштабы накрутки, т.е. долю фальшивых отзывов?

Решение. *Пункт 1. Метрики для определения успешности проекта*

Прежде чем отвечать на поставленный вопрос, определим цели проекта и разберёмся, что можно назвать хорошим результатом. К чему мы хотим прийти?

(1) Пользователи доверяют оценке продавца (пользователи охотнее покупают товары у продавцов с более высоким рейтингом);

(2) Продавцы с высоким рейтингом действительно добросовестны (среди продавцов с высоким рейтингом минимален процент неудачных покупок)

(3) Пользователи мотивированы оставлять отзывы о совершенных покупках (процент покупок, после которых следуют отзывы, со временем увеличивается)

Как можно отследить выполняемость этих пунктов?

Используемые метрики:

(1) Распределение количества продаж в зависимости от рейтинга продавца (хорошо, если с ростом рейтинга растёт количество продаж)

(2) Процент верно оценённых фейковых отзывов (возьмём тестовую выборку, для которой вручную оценим представленные отзывы (предполагаем, что человек с большой точностью сможет распознать фейковый отзыв) и будем тестировать нашу модель;

(3) Количество отзывов, которым предшествует поиск и долгое блуждание залогиненного пользователя по сайту (отследим, что человек действительно выбирал товар и рассматривал нескольких продавцов)

Пункт 2. Как оценить масштабы накрутки

Подумаем, как можно определить, фейковый отзыв или настоящий. Воспользуемся тем, что мы располагаем информацией об истории поиска пользователя и других показателях его деятельности на Авито: мы знаем (предположительно), сколько времени пользователь проводил на каждой странице, какие телефоны просматривал, открывал ли уже вкладку отзывы и т.д.

Сперва попробуем отсечь фейковые отзывы только по предыстории пользователя, не рассматривая их содержание.

Будем считать отзыв настоящим, если до написания отзыва $(*)$:

- пользователь искал товары, совпадающие по названию с товарами продавца;
- пользователь провёл на странице продавца больше 1 минуты (время опционально, но можно начать с этого; логика следующая: если человек собирается совершить покупку у продавца, он хотя бы минимально изучит его профиль);
- пользователь открывал фото товара (опять же, предполагаем, что если пользователь планирует совершить покупку, он хотя бы минимально изучит товар прежде чем связаться с продавцом);

- пользователь открывал отзывы на продавца прежде чем смотреть телефон или связываться с продавцом в чате;
- пользователь смотрел телефон или связывался с продавцом в чате.

Будем считать, что этот набор условий отсечёт фейковые отзывы. Однако, наравне с ними он также отсечёт и часть настоящих, поскольку некоторые отзывы оставляются с других девайсов (это проблема, если пользователь не залогинен) или спустя долгое время после покупки.

Чтобы не потерять слишком много настоящих отзывов, но учесть первичные рассуждения о правдивости отзыва и истории пользователя, введём коэффициент значимости k, который будет тем больше, чем больше пунктов из $(*)$ выполнено. Будем включать рейтинг i-го отзыва в итоговый рейтинг продавца с весом k_i.

Кроме того, назовём фейковыми отзывы:

- которые дублируются с другими (например, один продавец решил закинуть сразу нескольким своим конкурентам одинаковые плохие отзывы); Однако, у этого пункта есть явные недостатки – как минимум, неудобство хранения и обработки всех отзывов; однако если эта проблема будет решена, почему бы и не учесть такой показатель :) Например, можно закреплять за залогиненным пользователем
- если несколько отзывов отправлены подряд с небольшим временным интервалом (в пределах 5 минут, например) и обладают одинаковой оценкой.