Графы в python

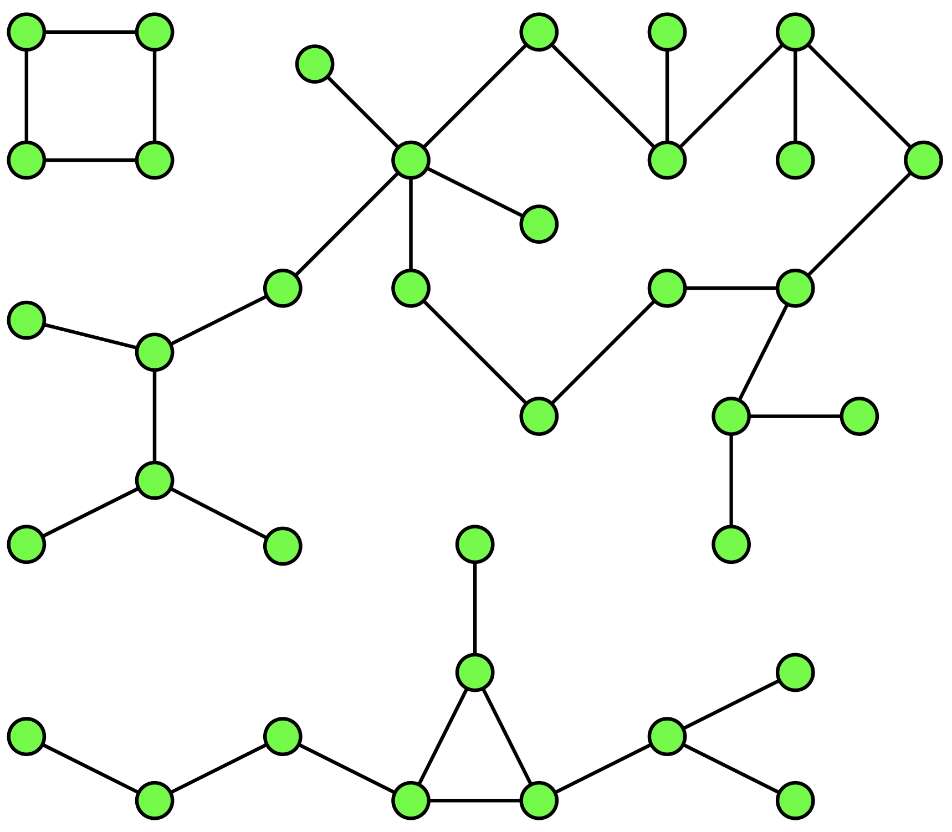
Граф состоит из отдельных точек (вершин) и соединяющих их линий (ребер). Вершины могут обозначать любые объекты: пользователей сайта, компьютеры корпоративной сети, населенные пункты на карте. Ребра отображают связи между этими объектами: каких пользователей тот или иной человек занес в «друзья», какие компьютеры объединяются прямым подключением, между какими городами проложены дороги. Как нетрудно догадаться, такая схема позволяет решать множество прикладных задач, и для каждой из них существует множество подходящих графов.

Такие структуры гораздо сложнее привычных таблиц баз данных. Представьте себе, что вам нужно создать карту значимых объектов обычного городского района, чтобы понять, где чаще всего бывают его жители. Вам нужно отметить все магазины, школы, МФЦ, и определить вес каждой из точки. Такой код непросто написать с нуля, а полученный результат очень сложно визуализировать в пригодном для работы формате — хаос повседневности формирует очень запутанные связи, даже если мы говорим про район, а не про город или страну.

К счастью, Python-программистам не приходится заниматься подготовительной работой. Они могут воспользоваться специальной библиотекой, которая объединяет все необходимые инструменты для создания и анализа графов — NetworkX.

**1. Разбивка вершин на группы**

Анализ связных компонентов (connected components) позволяет вам определить внутренние множества внутри вашего графа. Например, вы можете разбить посетителей интернет-магазина на группы по их предпочтениям или географической принадлежности. Или связать между собой подозрительные транзакции в банковской системе, чтобы предположить, что все эти действия провела одна группа мошенников.



**2. Поиск кратчайшего пути**

Вычисление наименьшего расстояния между двумя точками — одна из старейших задач программирования. Один из наилучших алгоритмов для ее решения еще в конце 50-х разработал нидерландский математик Эдсгер Дейкстра (Edsger Dijkstra), причем человечество может благодарить за это открытие… его невесту.

Как рассказывал сам Дейкстра, однажды его так утомили походы по магазинам вместе с молодой женой, что во время краткой передышки он задумался — как оптимизировать этот процесс? Буквально за 20 минут он разработал метод, который теперь называется [алгоритмом Дейкстры](https://ru.wikipedia.org/wiki/%25D0%2590%25D0%25BB%25D0%25B3%25D0%25BE%25D1%2580%25D0%25B8%25D1%2582%25D0%25BC_%25D0%2594%25D0%25B5%25D0%25B9%25D0%25BA%25D1%2581%25D1%2582%25D1%2580%25D1%258B).

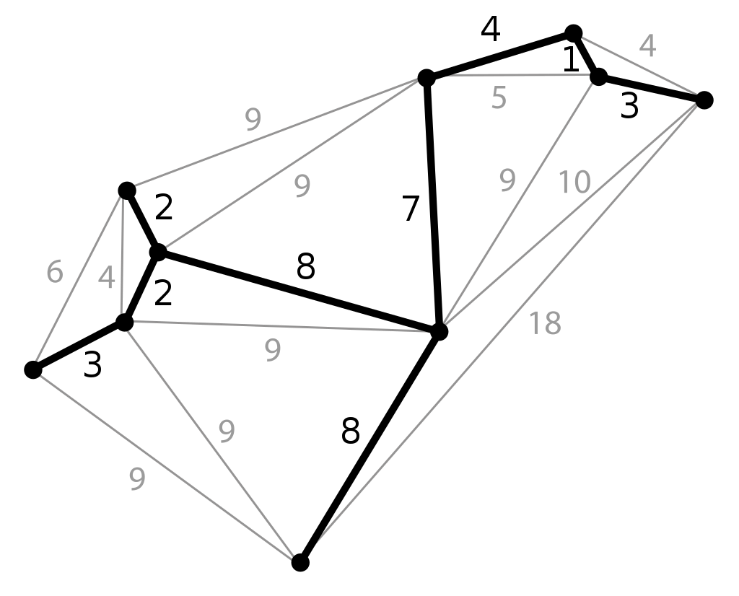
Метод прост, как все гениальное. Для начала Data Scientist узнает расстояния от начальной точки до каждой точки в пути. Эти значения присваиваются граням графа. Далее алгоритм поочередно проходит по каждой вершине, на каждом шаге определяя наименьшее расстояние до очередной точки. Когда все точки посещены, вы получаете наименьший маршрут.



**3 Поиск оптимального пути**

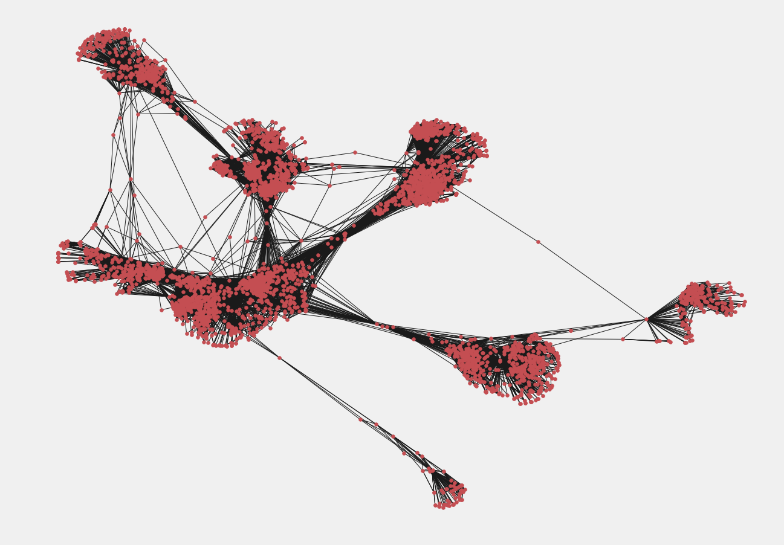
Всем знакомы головоломки, в которых нужно начертить какую-либо фигуру, не проходя дважды через одну точку. В профессии Data Scientist такие задачи возникают, когда нужно проложить оптимальный путь для кабеля, помочь торговым представителям или мастерам по обслуживанию банкоматов спланировать маршрут, и так далее.

В каждом случае помогает алгоритм минимального остовного дерева (Minimum Spanning Tree). Этот термин обозначает подграф, который объединяет ребра с минимально возможной суммой весов. Это не всегда означает, что одна вершина не может иметь больше двух ребер — главное значение имеет итоговая сумма их значений.



**4. Рейтинг вершин графа**

Еще одна нередкая задача — найти, какие вершины имеют самое важное значение в рамках того или иного явления. Например, поисковики определяют вес разных интернет-страниц, академики считают, какие научные работы чаще всего ссылаются исследователи, социальные сети отслеживают пользователей с наибольшим количеством подписок.



**5. Центральность вершин**

Понятие центральности обозначает расстояние какой-либо вершины графа до его центра. Измерение центральности — это еще один способ определить влияние разных участников базы данных.

В отличие от предыдущего метода, определение центральности направлено на выявление не самой главной вершины, а локальных «лидеров», которых может быть и пять, и десять. Таким образом Data Scientist может изучать распространение информации в сообществе (определять изначальное сообщение и следить за его распространением через несколько точек входа), отслеживать развитие болезней, контролировать телекоммуникационные и прочие сети.

Существует множество метрик центральности. В нашем примере мы расскажем, как найти на графе вершины, которые чаще всего связывают между собой разрозненные группы точек. На практике это может быть перевалочный пункт на географическом маршруте, «бутылочное горлышко» в каких-либо процессах, человек-посредник, который знакомит между собой других людей. В теории графов это и называется степенью посредничества (Betweenness Centrality).

