Задание по программированию: Размещение баннеров

|  |  |
| --- | --- |
| **Срок сдачи** | Задание необходимо было сдать апрель 16, 11:59 вечера PDT  Вы все еще можете выполнить задание до завершения курса. |

Представим, что международное круизное агентство *Carnival Cruise Line* решило себя разрекламировать с помощью баннеров и обратилось для этого к вам. Чтобы протестировать, велика ли от таких баннеров польза, их будет размещено всего 20 штук по всему миру. Вам надо выбрать 20 таких локаций для размещения, чтобы польза была большой и агентство продолжило с вами сотрудничать.

Агентство крупное, и у него есть несколько офисов по всему миру. Вблизи этих офисов оно и хочет разместить баннеры — легче договариваться и проверять результат. Также эти места должны быть популярны среди туристов.

Для поиска оптимальных мест воспользуемся базой данных крупнейшей социальной сети, основанной на локациях — Foursquare.

Часть открытых данных есть, например, на сайте archive.org:

<https://archive.org/details/201309_foursquare_dataset_umn>

Скачаем любым удобным образом архив fsq.zip с этой страницы.

Нас будет интересовать файл checkins.dat. Открыв его, увидим следующую структуру:

id | user\_id | venue\_id | latitude | longitude | created\_at

---------+---------+----------+-------------------+-------------------+---------

------------

984301 | 2041916 | 5222 | | | 2012-04-21 17:39:01

984222 | 15824 | 5222 | 38.8951118 | -77.0363658 | 2012-04-21 17:43:47

984315 | 1764391 | 5222 | | | 2012-04-21 17:37:18

984234 | 44652 | 5222 | 33.800745 | -84.41052 | 2012-04-21 17:43:43

Для удобной работы с этим документом преобразуем его к формату csv, удалив строки, не содержащие координат — они неинформативны для нас:

id,user\_id,venue\_id,latitude,longitude,created\_at

984222,15824,5222,38.8951118,-77.0363658,2012-04-21T17:43:47

984234,44652,5222,33.800745,-84.41052,2012-04-21T17:43:43

984291,105054,5222,45.5234515,-122.6762071,2012-04-21T17:39:22

...

С помощью pandas построим DataFrame и убедимся, что все 396634 строки с координатами считаны успешно.

Теперь необходимо кластеризовать данные координаты, чтобы выявить центры скоплений туристов. Поскольку баннеры имеют сравнительно небольшую площадь действия, нам нужен алгоритм, позволяющий ограничить размер кластера и не зависящий от количества кластеров.

Эта задача — хороший повод познакомиться с алгоритмом MeanShift, который мы обошли стороной в основной части лекций. Его описание при желании можно посмотреть в [sklearn user guide](http://scikit-learn.org/stable/modules/clustering.html" \l "mean-shift" \t "_blank), а чуть позже появится дополнительное видео с обзором этого и некоторых других алгоритмов кластеризации. Используйте MeanShift, указав bandwidth=0.1, что в переводе из градусов в метры колеблется примерно от 5 до 10 км в средних широтах.

**Примечание:**на 396634 строках кластеризация будет работать долго. Быть очень терпеливым не возбраняется — результат от этого только улучшится. Но для того, чтобы сдать задание, понадобится сабсет из первых 100 тысяч строк. Это компромисс между качеством и затраченным временем. Обучение алгоритма на всём датасете занимает около часа, а на 100 тыс. строк — примерно 2 минуты, однако этого достаточно для получения корректных результатов.

Некоторые из получившихся кластеров содержат слишком мало точек — такие кластеры не интересны рекламодателям. Поэтому надо определить, какие из кластеров содержат, скажем, больше 15 элементов. Центры этих кластеров и являются оптимальными для размещения.

При желании увидеть получившиеся результаты на карте можно передать центры получившихся кластеров в один из инструментов визуализации. Например, сайт [mapcustomizer.com](https://www.mapcustomizer.com/" \t "_blank) имеет функцию Bulk Entry, куда можно вставить центры полученных кластеров в формате:

38.8951118,-77.0363658

33.800745,-84.41052

45.5234515,-122.6762071

...

Как мы помним, 20 баннеров надо разместить близ офисов компании. Найдем на Google Maps по запросу *Carnival Cruise Line* адреса всех офисов:

33.751277, -118.188740 (Los Angeles)

25.867736, -80.324116 (Miami)

51.503016, -0.075479 (London)

52.378894, 4.885084 (Amsterdam)

39.366487, 117.036146 (Beijing)

-33.868457, 151.205134 (Sydney)

Осталось определить 20 ближайших к ним центров кластеров. Т.е. посчитать дистанцию до ближайшего офиса для каждой точки и выбрать 20 с наименьшим значением.

**Примечание:** при подсчете расстояний и в кластеризации можно пренебречь тем, что Земля круглая, так как в точках, расположенных близко друг к другу погрешность мала, а в остальных точках значение достаточно велико.

Для сдачи задания выберите из получившихся 20 центров тот, который наименее удален от ближайшего к нему офиса. Ответ в этом задании — широта и долгота этого центра, записанные через пробел.