# Постановка задачи.

В наше время существует очень много различных музыкальных сервисов, как, например, Яндекс.музыка. Такие сервисы ориентированы на потребителей, поэтому очень важно знать, какую музыку предпочитает тот или иной пользователь, чтобы предложить ему то, что ему понравится. Например, человеку слушает преимущественно классическую музыку, вряд ли стоит предлагать ему рэп. И тогда владельцами сервисов встаёт задача сформировать максимально точные рекомендации, чтобы люди продолжали пользоваться их сервисом. Для этого нужно определить музыку, которая понравилась бы пользователю, была бы близка по жанру, стилю, ведущему инструменту к тому, что уже определено как любимая музыка.

В данной работе я буду рассматривать обнаружение описательных музыкальных жанров с помощью кластеризации K-средних (K-means clustering)

# Теоретическая часть

Кластерный анализ — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Кластерный анализ — это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы (кластеры). Кластер — группа элементов, характеризуемых общим свойством, главная цель кластерного анализа — нахождение групп схожих объектов в выборке. Спектр применений кластерного анализа очень ВЕСТНИК МГТУ МИРЭА HERALD of MSTU MIREA 2014 № 2 июнь 2014 выпуск 3 87 широк: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, дистанционном зондировании и других дисциплинах.

Кластерный анализ выполняет следующие основные задачи:

* Разработка типологии или классификации.
* Исследование полезных концептуальных схем группирования объектов.
* Порождение гипотез на основе исследования данных.
* Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных. Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:
* Отбор выборки для кластеризации.
* Определение множества переменных, по которым будут оцениваться объекты в выборке.
* Вычисление значений той или иной меры сходства между объектами.
* Применение метода кластерного анализа для создания групп сходных объектов.
* Проверка достоверности результатов кластерного решения.

Кластерный анализ предъявляет следующие требования к данным: во-первых, показатели не должны коррелировать между собой; во-вторых, показатели должны быть безразмерными; в-третьих, их распределение должно быть близко к нормальному; в-четвёртых, показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов; в-пятых, выборка должна быть однородна, не содержать «выбросов». Если кластерному анализу предшествует факторный анализ, то выборка не нуждается в «ремонте» — изложенные требования выполняются автоматически самой процедурой факторного моделирования. В противном случае выборку нужно корректировать.

## Методы кластерного анализа

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд групп подходов (некоторые методы можно отнести сразу к нескольким группам и потому предлагается рассматривать данную типизацию как некоторое приближение к реальной классификации методов кластеризации).

Вероятностный подход. Предполагается, что каждый рассматриваемый объект относится к одному из классов. Некоторые авторы (например, А. И. Орлов) [5] считают, что данная группа вовсе не относится к кластеризации и противопоставляют ВЕСТНИК МГТУ МИРЭА HERALD of MSTU MIREA 2014 № 2 июнь 2014 выпуск 3 88 её под названием «дискриминация», то есть выбор отнесения объектов к одной из известных групп (обучающих выборок).

## Метод k-средних

Наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом.

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:



где k — число кластеров, Si— полученные кластеры,  — центры масс векторов .

Алгоритм представляет собой версию EM-алгоритма, применяемого также для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

Недостатки:

* Не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов.
* Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
* Число кластеров надо знать заранее.

Входными данными в методе k-средних является только матрица наших X. Как правило, мы формируем её так, чтобы каждая строка представляла отдельный пример (образец), а каждый столбец – отдельный признак или, пользуясь терминами из статистики, фактор. Обычно мы говорим, что есть N примеров и D признаков, так что X представляет собой матрицу размерности NxD.

В алгоритме метода k-средних есть два основных этапа. Сначала мы выбираем k разных центров кластеров – как правило, это просто случайные точки в наборе данных. Затем мы переходим к нашему основному циклу, который также состоит из двух этапов. Первый – это выбор, к какому из кластеров принадлежит каждая точка из X. Для этого мы берём каждый пример и выбираем кластер, чей центр ближе всего. Не забывайте, что вначале мы выбираем центры случайным образом. Второй этап – заново вычислить каждый центр кластера, основываясь на множестве точек, которые к нему приписаны. Для этого берутся все соответствующие примеры и вычисляется их среднее значение, отсюда и название метода – «метод k-средних». Всё это делается до тех пор, пока алгоритм не сойдётся, то есть пока не прекратится изменение в распределении точек по кластерам или в координатах центров кластеров. Как правило, это происходит очень быстро – в районе от 5 до 15 проходов цикла. Это сильно отличается от градиентного спуска в глубоком обучении, где могут пройти тысячи итераций, пока не произойдёт схождение.

# Dataset

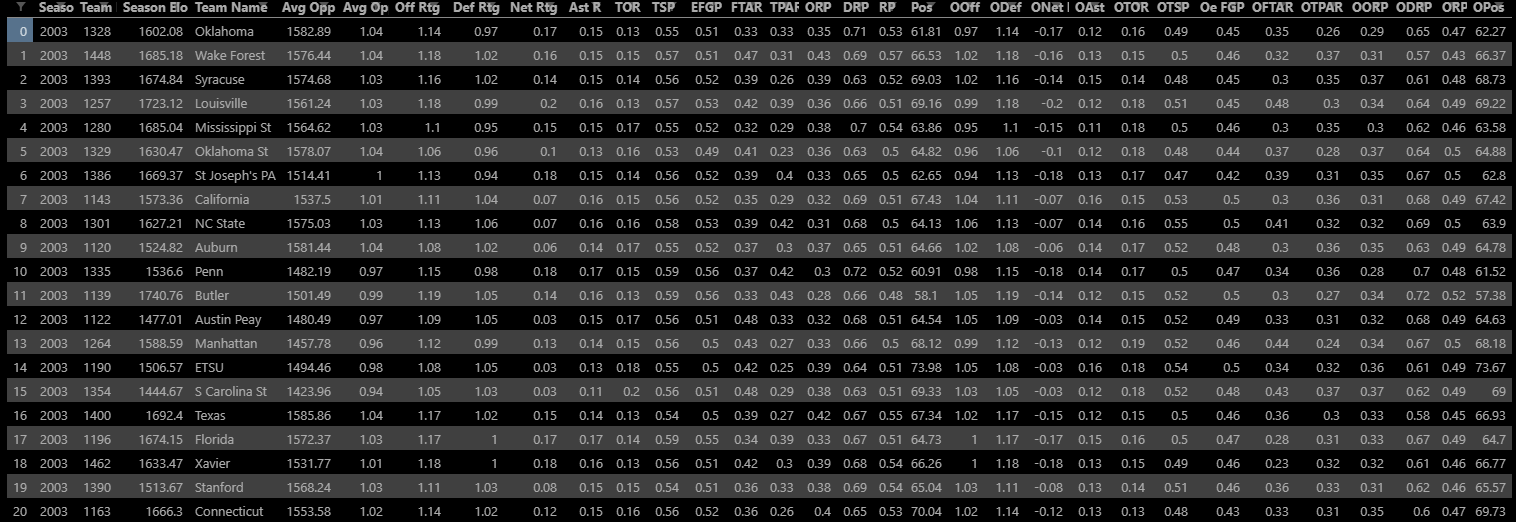


Рисунок Первые 20 строк набор данных, состоящих из 1064 записей

# Исходный код

См. в приложении.

# Результаты

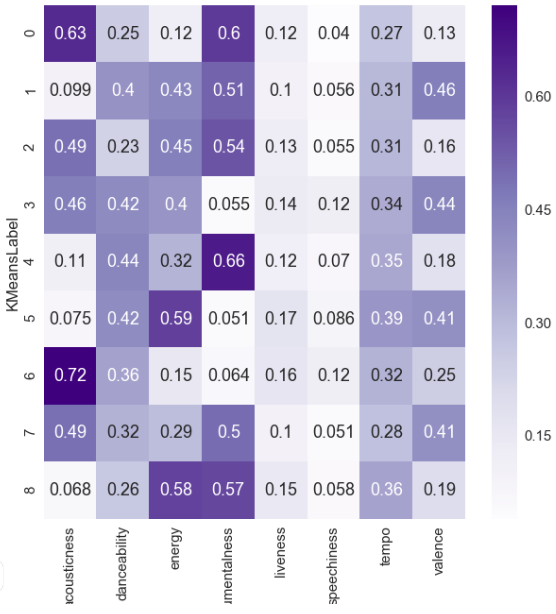


Рисунок Тепловая карта значений звуковых объектов по метке K-Means для всей музыки.

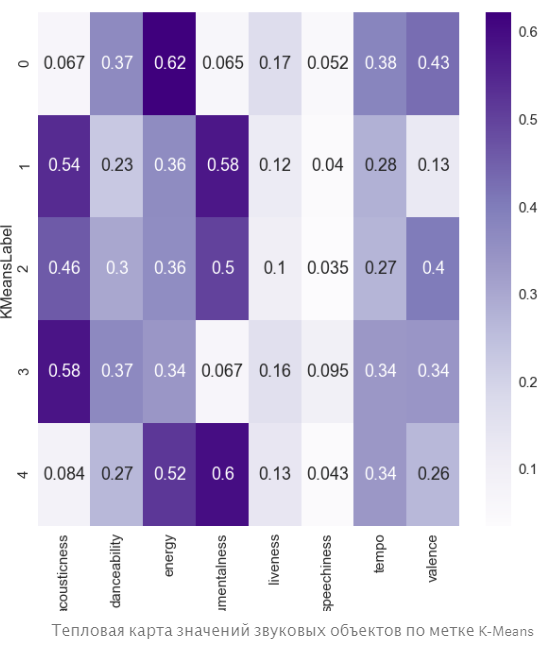
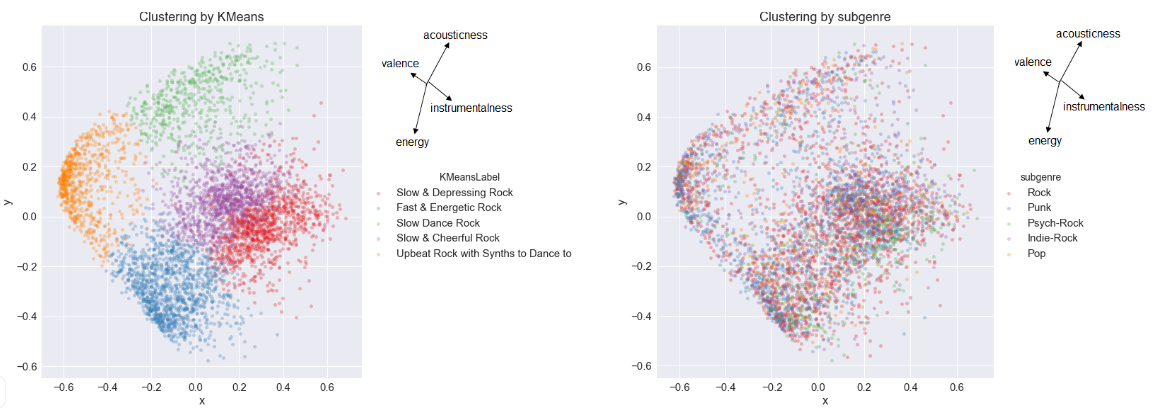
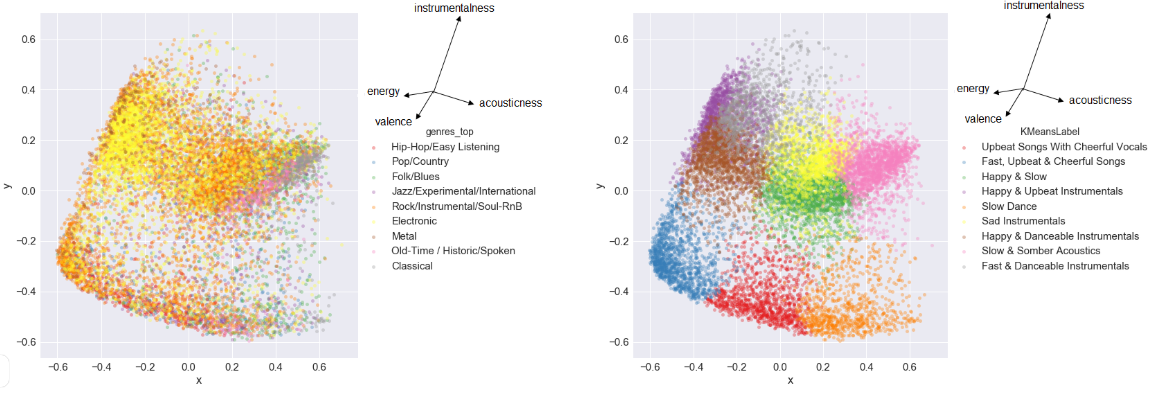


Рисунок Тепловая карта значений звуковых объектов по метке K-Means для рок-музыки



# Выводы

Кластеризация музыки в жанры на основе звуковых функций позволяет описывать музыку по-новому. Объединяя эти жанры с уже используемыми метками, можно получить новые и более точные жанры. Таким образом, можно с уверенностью заявить, что благодаря рассмотренному выше методу можно формировать очень точные рекомендации музыки.