
hadoop

Содержание

1. *Постановка задачи*

2. *Подготовка данных и первичный анализ*
   - *Подготовка обучающих данных*
   - *Визуализация данных*
   - *Использование Count Vectorizer*

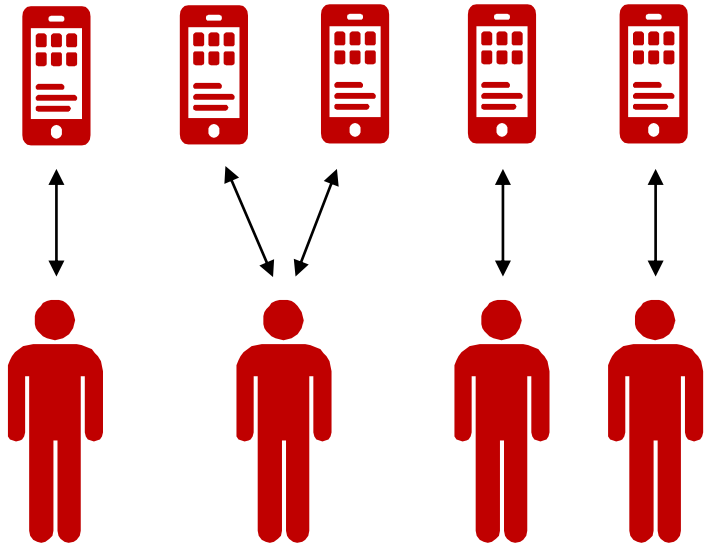
3. *Анализ, построение и оценка алгоритма*
   - *Анализ векторов*
   - *Аггломеративный алгоритм*
   - *Задание расстояния*
   - *Метрика качества*

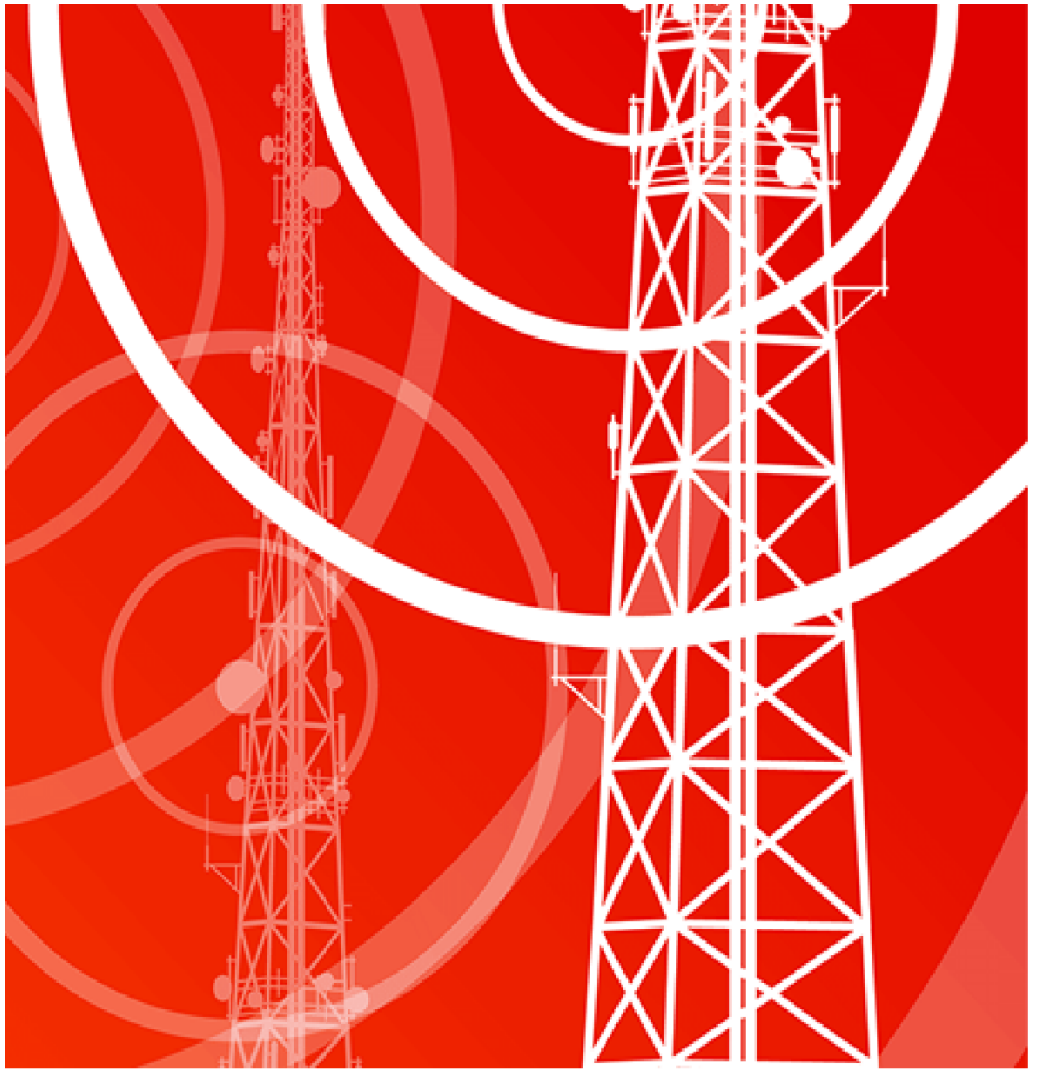
4. *Вывод*

Введение и постановка задачи

Постановка задачи

Большинство людей используют несколько сим-карт. Компания сталкивается с задачей выявления абонентов, которые являются одним человеком



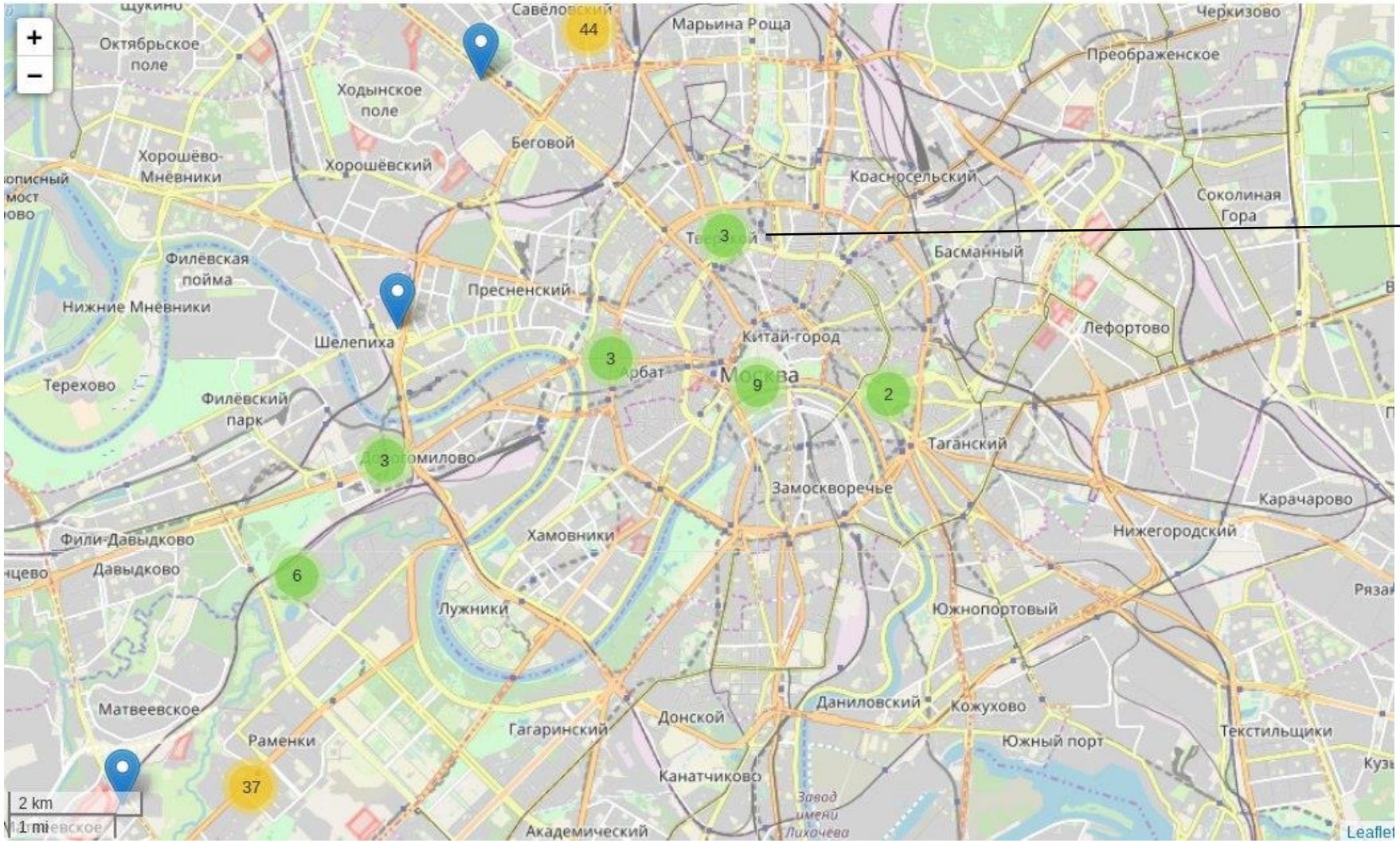


Подготовка данных и Первичный анализ

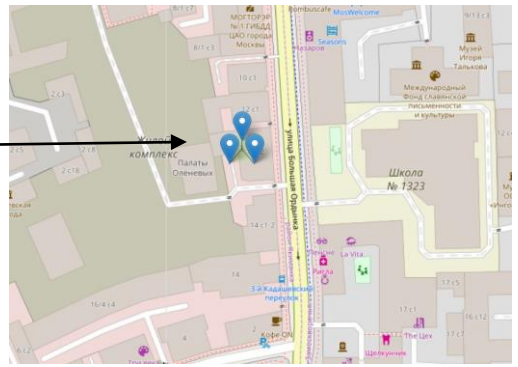
Подготовка обучающих данных

Для последующей проверки алгоритма разделим данные Train (Facts.csv) на три части. В каждой из них есть номера-дубликаты. Протестируем алгоритм на них чтобы выявить параметры для обработки validate части.

Данные по маршруту одного номера



Некоторые точки представляют группы



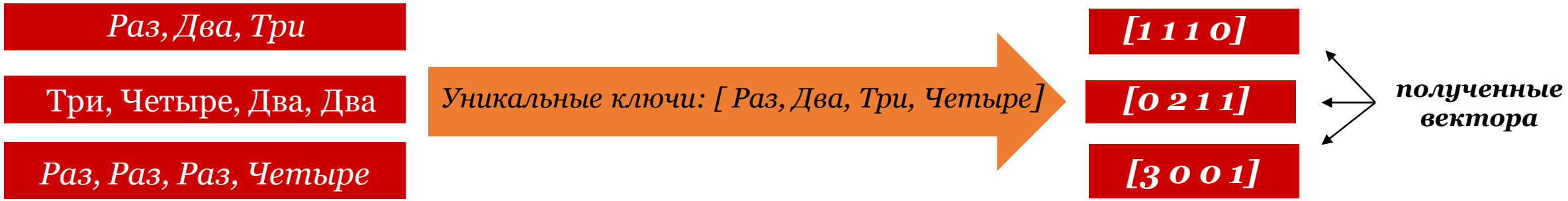
Нужно представить данные о маршрутах номера в виде вектора

В этом поможет метод Count Vectorizer

Count Vectorizer

Задача выявления дубликатов номеров похожа на задачу выявления дубликатов текста.
Данные (вектора) в решениях таких задач представляются с помощью алгоритма Count Vectorizer

Принцип работы метода:



Каждый текст представляется в виде вектора, где координата, это число вхождений того или иного уникального слова

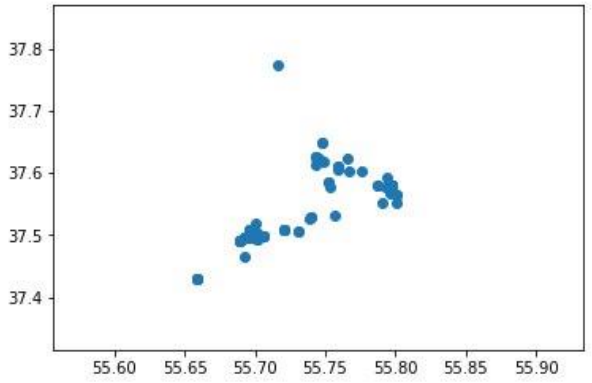
Count Vectorizer

На этих графиках нарисованы точки (координаты вышек) в которых были отмечены два номера, соответствующих одному человеку.

Тогда ключевое предположение алгоритма будет заключаться в том, что номера принадлежат одному человеку если у этих номеров много общих совпадающих вышек.

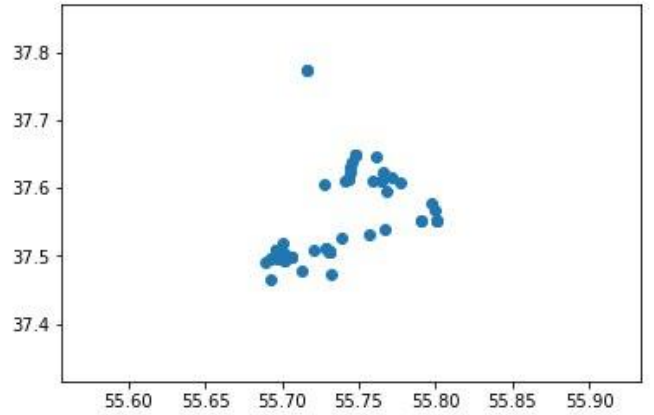
Представим данные таким образом, чтобы каждому уникальному номеру соответствовала последовательность идентификаторов вышек в которых он был (заметим, что координаты вышек и идентификаторы вышек взаимно-однозначны).

Номер: 158510090027



Координаты: Широта и Долгота

Номер: 15852885087



номер	***Последовательность идентификаторов вышек (Lac и CID)***	**Count Vectorizer**	***вектор***
158510090027	*778248428 775230479 774226211 …*	**Count Vectorizer**	*[1 0 3 2 1 …]*

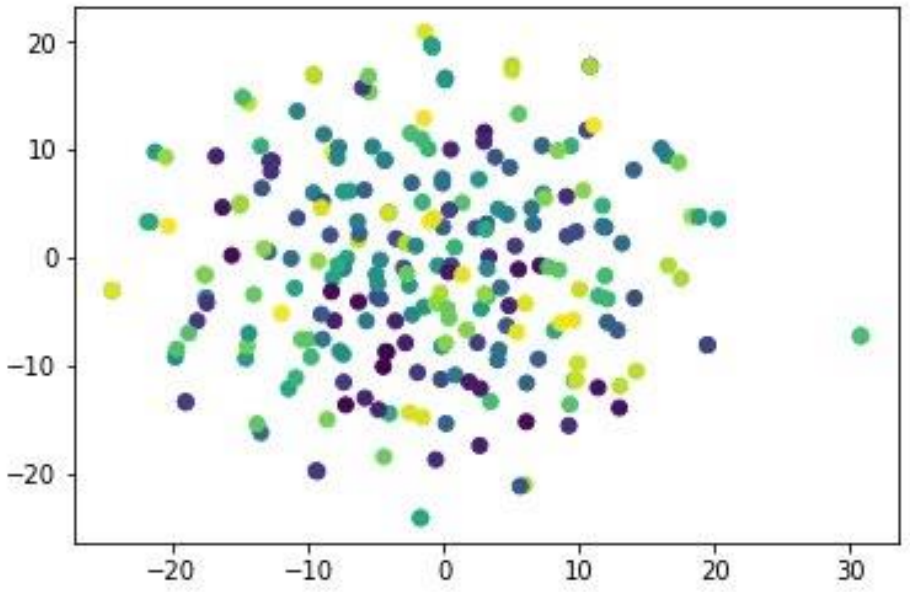
Анализ, построение и оценка алгоритма

Визуализация и анализ

Для начала стоит визуализировать полученные после Count Vectorizer вектора.

На картинке представлены все соответствующие номерам (векторам) точки в двумерном случае.*

Точки одного цвета принадлежат одному человеку, как видно из картинки большинство из них располагаются парами.



*для визуализации я понизил размерность векторов полученных методом Count Vectorizer до двух используя алгоритм визуализации T-SNE

Отсюда умозаключение о том, что ближайшие точки хорошо бы объединить в один кластер, а остальные оставить не тронутыми.

Подойдет аггломеративный алгоритм кластеризации.

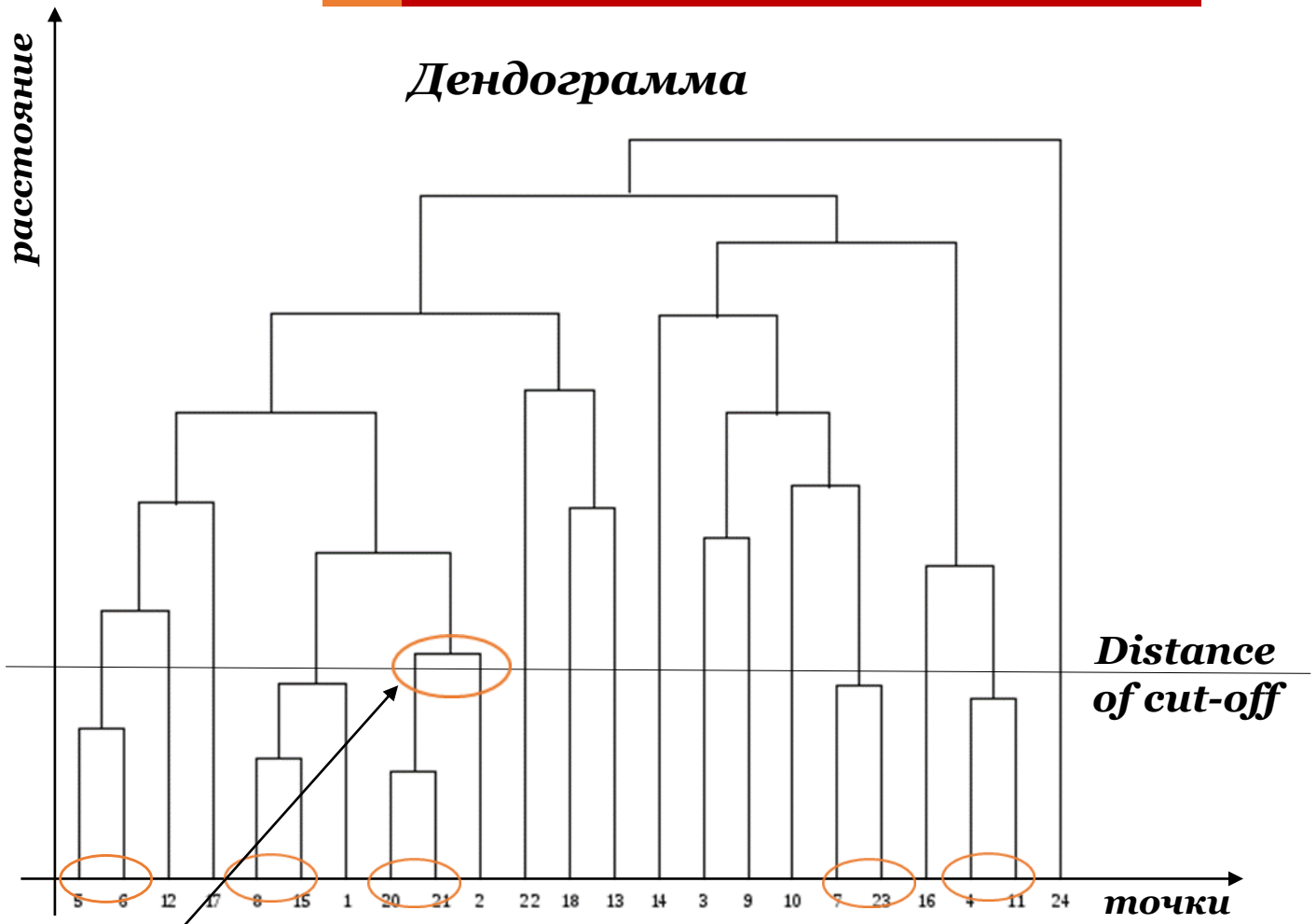
Агломеративный алгоритм

Агломеративный алгоритм объединяет близлежащие точки в один кластер.

Иными словами, если маршруты двух номеров похожи между собой (расстояние между соответствующими векторами маленькое), то мы объединим их в один кластер.

Однако в таком подходе есть две проблемы:

- ***Как задать расстояние***
- ***При последовательном объединении номеров в один кластер могут попасть номера похожие, но не соответствующие одному абоненту. Как найти подходящий момент чтобы обрезать ветвь***



В этот момент в кластер попадает лишний номер, нужно обрезать дендограмму до этого момента

Точки в оранжевых кружках будут определены как дубликаты

Расстояние

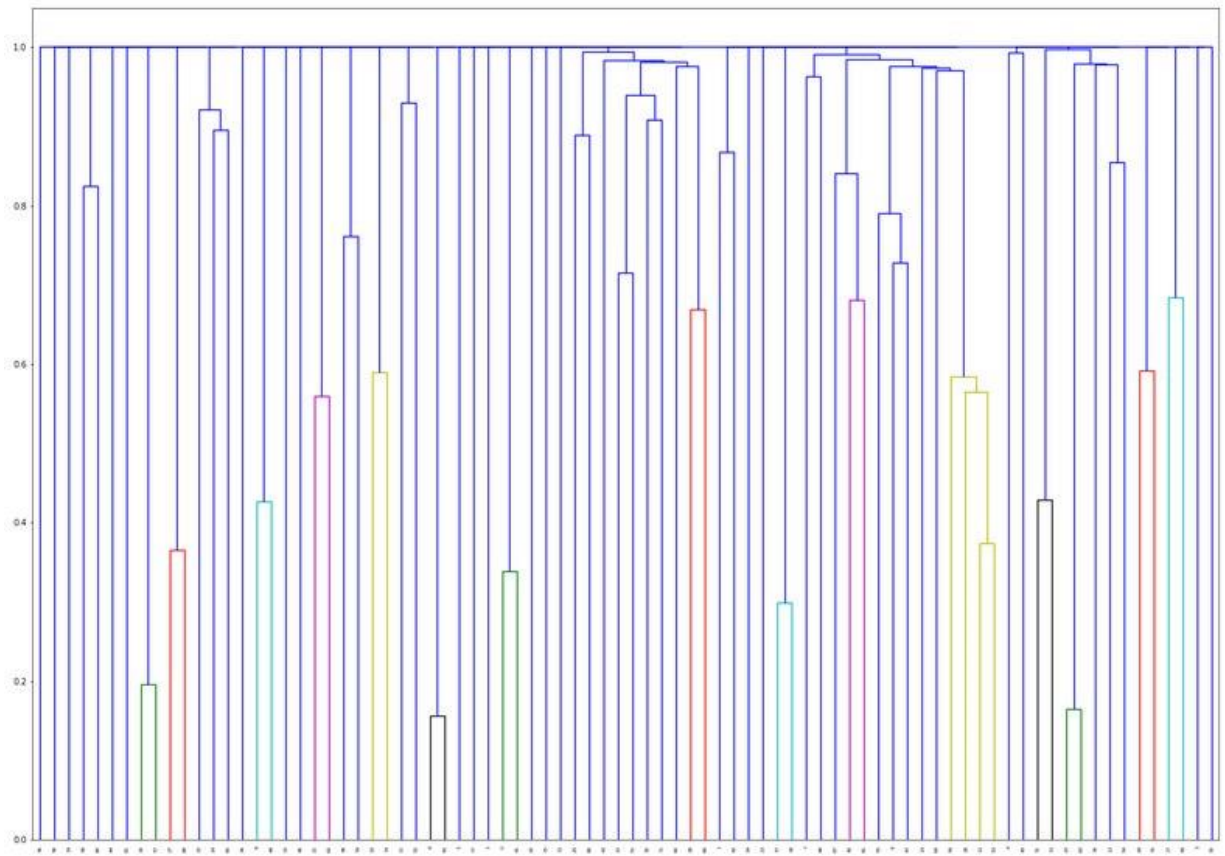
В качестве расстояния между векторами, полученными с помощью Count Vectorizer, буду использовать коэффициент корреляции Пирсона:

$$r_{xy} = \frac{\sum_{i=1}^{i=N} (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{X})^2 \cdot \sum_i (y_i - \overline{Y})^2}}$$

На полученной с использованием коэффициента Пирсона в качестве расстояния дендограмме видно, что ближайшие точки (номера) попали в один класс, при этом номера почти не попадают в чужие классы.

Таким образом использование в качестве расстояния коэффициента корреляции позволяет справиться с проблемой попадания в кластеры лишних номеров.

Осталось подобрать Distance of cut-off



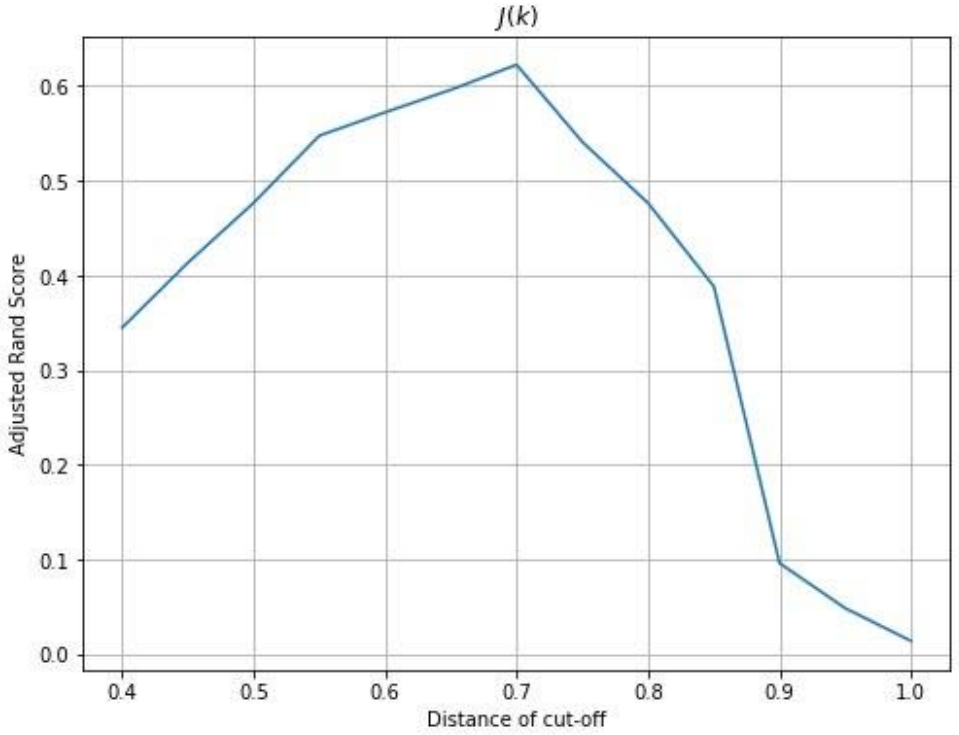
Метрика качества

Для подбора оптимального расстояния и других параметров нужно оценивать насколько хорошо работает алгоритм, для этого нужна метрика качества.

Так как обучающие данные представляют пары номеров, то в нашем случае хорошо подойдет Adjusted Rand Index.

$$\mathrm{Rand}(\alpha, \beta) = \frac{a + d}{a + b + c + d}$$

где:
- *a количество пар объектов, находящихся в одинаковых кластерах α и β*
- *b и c количество пар объектов в одном и том же кластере в α (β), но в разных α (β)*
- *d количество пар объектов в разных кластерах в α и β*



Adjusted Rand Index – корректировка Rand Index:

$$\mathrm{ARI}(\alpha, \beta) = \frac{\mathrm{Rand}(\alpha,\beta) - Expected}{Max - Expected}$$

В этом примере оптимальное значение ARI достигается при Distance cut-off равном 0.7

Вывод

Таким образом алгоритм достаточно неплохо выделяет дубликаты из массы маршрутов номеров, однако основная проблема заключается в подборе расстояния на котором нужно обрезать дендограмму. Чем оно больше, тем больше дубликатов найдет алгоритм, однако возможен риск попадания в один кластер номеров, не принадлежащих одному человеку номеров.

Спасибо за внимание

Шевцов Антон
Shevan05@gmail.com
8(919)-723-72-44