КЛАСТЕРИЗАЦИЯ ТЕКСТОВ

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание

1. Основные определения
2. Автоматическая кластеризация текстов:
   - постановка задачи
   - виды алгоритмов кластеризации
   - алгоритмы и примеры применения
3. Оценка качества кластеризации
4. Домашнее задание

Вопрос

В чем основные отличия кластеризации от классификации?

Основные определения

Классификация (или **рубрицирование**) – отнесение объекта к заранее известным классам (рубрикам)

- классы: с заданными характеристиками, иерархическая система классов
- обычно классифицируют по «содержанию» объектов

Кластеризация – разбиение заданного множества объектов на подмножества (кластеры)

- объекты в кластерах похожи по смыслу/теме/структуре/…
- характеристики, количество, структура кластеров заранее не заданы

Вопрос

Какие цели может преследовать кластеризация?

Цели кластеризации

- Понять структуру множества объектов, разбив его на группы схожих объектов

Пример: в маркетинге, выделяют отдельные группы клиентов/покупателей/товаров и разрабатывая для каждой отдельную стратегию

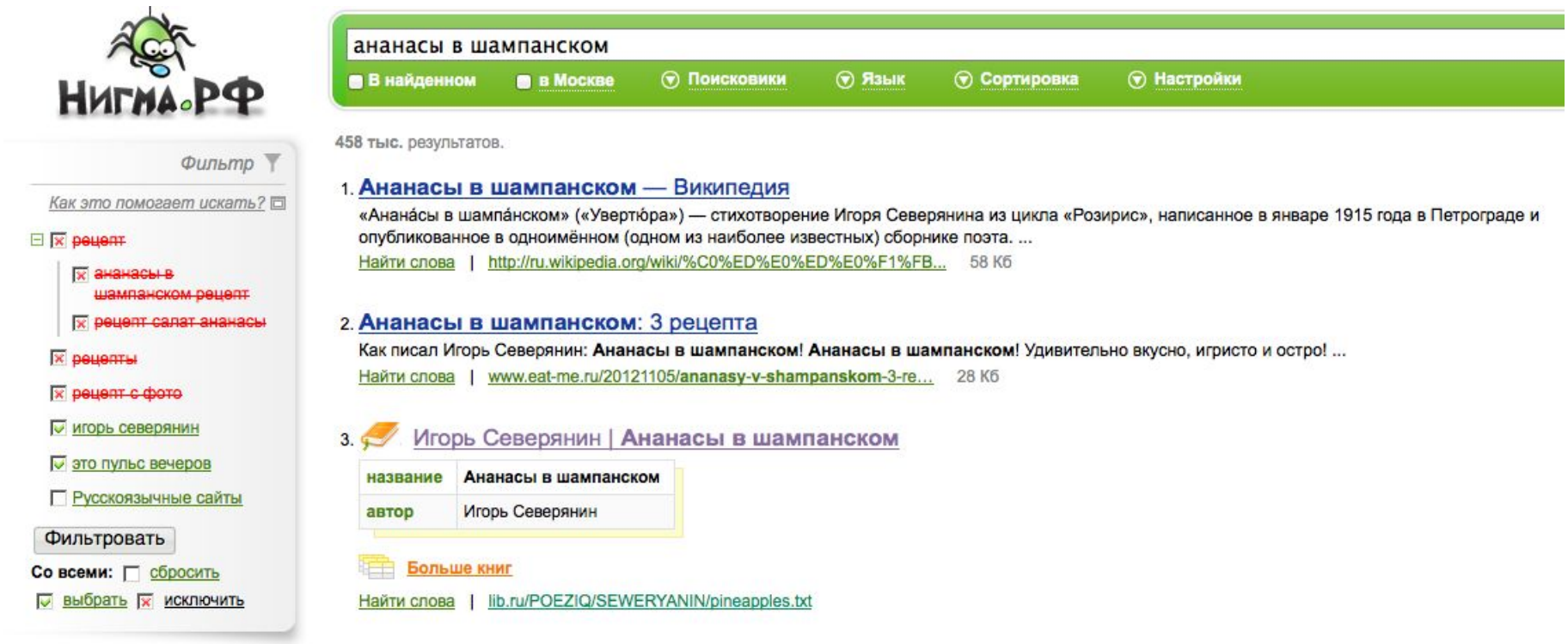
- Сократить объём хранимых данных, оставив по одному наиболее типичному представителю от каждого кластера

Пример: …

- Выделить нетипичные объекты, которые не подходят ни к одному из кластеров

Пример: …

Пример: «интеллектуальная» группировка результатов при информационном поиске



Сейчас кластеризация часто – один из этапов анализа данных

Формальная постановка задачи автоматической кластеризации

- Имеется множество объектов
  $D = \{d_1, \ldots, d_{|D|}\}$
- Существует множество «тематических классов»
  $C = \{c_1, \ldots, c_{|C|}\}$
- Предполагается, что можно автоматически разбить *D* на кластеры и они будут соответствовать *C*
- Задача сводится к поиску такого *C*, которое являлось бы оптимальным в соответствии с некоторым критерием качества
- ✔ Нужно определить <u>критерий качества разбиения</u>

Какими должны быть кластеры?

Внутри каждого кластера должны оказаться «похожие» объекты, а объекты разных кластеров должны быть «не похожи»

Другими словами:

- схожесть между объектами из одного кластера должна быть как можно больше
- схожесть между объектами из разных кластеров должна быть как можно меньше

- ✔ Нужно определить критерий схожести между объектами
- ✔ Алгоритм должен самостоятельно принимать решение о количестве и составе кластеров

Кластеризация текстов (документов)

- Документов представляются как вектора в пространстве признаков

$$d_i = (d_{i1}, \ldots, d_{i|T|}),$$ где

d_{ij} – вес j-ого признака в i-ом документе $0 \le d_{ij} \le 1$,
$|T|$ – количество различных признаков в D

- Для определения схожести документов обычно вычисляют расстояния между ними

Примеры: евклидова метрика, манхэттенское расстояние, косинусное расстояние

✔ Кто выбирает критерий схожести?

Примеры мер (1)

- Евклидово расстояние – геометрическое расстоянием в многомерном пространстве

$$\rho(x,y)=\sqrt{\sum_{i}^{n}(x_i-y_i)^2}$$

- Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам

$$\rho(x,y)=\sum_{i}^{n}(x_i-y_i)^2$$

- Манхэттенское расстояние (расстояние городских кварталов) – сумма разностей по координатам. Уменьшает влияние выбросов

$$\rho(x,y)=\sum_{i}^{n}|x_i-y_i|$$

Примеры мер (2)

- Расстояние Чебышева полезно для «различения» объектов, отличных в одной координате

$$\rho(x, y) = \max_i |x_i - y_i|$$

- Считающее расстояние – число координат, по которым векторы x и y различаются

$$\rho(x, y) = \sum_i^n [x_i \neq y_i]$$

- Косинусное расстояние

$$\rho(x, y) = \arccos\left(\frac{\sum_i^n x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}\right)$$

Всегда ли нужны именно подобные меры?

Задание 1

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

Построить для каждого текста представление в виде вектора. Вес: присутствует признак в документе или нет

Задание 1. Ответ

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

Построить для каждого текста представление в виде вектора. Вес: присутствует признак в документе или нет

w_1=(1, 1, 1, 1, 0, 0, 0, 0)
w_2=(1, 1, 1, 0, 1, 0, 0, 0)
w_3=(1, 1, 1, 1, 0, 0, 0, 0)
w_4=(0, 0, 0, 0, 0, 1, 1, 1)

Задание 2

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

Построить для каждого текста представление в виде вектора. Вес: *tf-idf*

Взгляд в прошлое

$$w_{ji} = f_{ji} \log(\frac{N}{N_i}) = tf_{ji} idf_i$$

	Карл	Клара	украсть	коралл	кларнет
df_i	3	3	3	2	1
idf_i	0	0	0	0,18	0,48

Документ 1			Документ 2			Документ 3		
слово	tf_{ji}	tf-idf	слово	tf_{ji}	tf-idf	слово	tf_{ji}	tf-idf
Карл	0,25	0	Карл	0,25	0	Карл	0,25	0
Клара	0,25	0	Клара	0,25	0	Клара	0,25	0
украсть	0,25	0	украсть	0,25	0	украсть	0,25	0
коралл	0,25	0,045	кларнет	0,25	0,12	коралл	0,25	0,045

Задание 2. Ответ

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

Построить для каждого текста представление в виде вектора. Вес: *tf-idf*

w1=(0.03, 0.03, 0.03, 0.075, 0, 0, 0, 0)
w2=(0.03, 0.03, 0.03, 0, 0.15, 0, 0, 0)
w3=(0.03, 0.03, 0.03, 0.075, 0, 0, 0, 0)
w4=(0, 0, 0, 0, 0, 0.2, 0.2, 0.2)

Задание 3

1. Вычислить *косинусное расстояние* для
$w_1=(1,1,1,1,0,0,0,0)$ и $w_2=(1,1,1,0,1,0,0,0)$
$w_1=(1,1,1,1,0,0,0,0)$ и $w_3=(1,1,1,1,0,0,0,0)$

2. Вычислить *евклидово расстояние* для
$w_3=(1,1,1,1,0,0,0,0)$ и $w_4=(0,0,0,0,0,1,1,1)$

$w_3=(0.03,0.03,0.03,0.075,0,0,0,0)$ и
$w_4=(0,0,0,0,0,0.2,0.2,0.2)$

3. Вычислить *манхэттенское расстояние* для
$w_1=(0.03,0.03,0.03,0.075,0,0,0,0)$ и
$w_4=(0,0,0,0,0,0.2,0.2,0.2)$

$w_2=(0.03,0.03,0.03,0,0.15,0,0,0)$ и
$w_3=(0.03,0.03,0.03,0.075,0,0,0,0)$

Задание 3. Обсуждение (1)

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

косинусное расстояние для

w_1 и w_2 ≈ $0{,}72w_1$ и w_3 = 0 w_1 и w_4 ≈ 1,57
w_2 и w_3 ≈ $0{,}72w_2$ и w_4 ≈ 1,57 w_3 и w_4 ≈ 1,57

евклидова метрика для

w_1 и w_2 ≈ $1{,}41w_1$ и w_3 = 0 w_1 и w_4 ≈ 2,65
w_2 и w_3 ≈ $1{,}41w_2$ и w_4 ≈ 2,65 w_3 и w_4 ≈ 2,65

манхэттенское расстояние для

w_1 и w_2 = 2 w_1 и w_3 = 0 w_1 и w_4 = 7
w_2 и w_3 = 2 w_2 и w_4 = 7 w_3 и w_4 = 7

Задание 3. Обсуждение (2)

1: *Карл у Клары украл кораллы*
2: *Клара у Карла украла кларнет*
3: *Клара у Карла украла кораллы*
4: *Простота – хуже воровства*

косинусное расстояние для

w_1 и w_2 ≈ $1{,}55w_1$ и w_3 = 0 w_1 и w_4 ≈ 1,57

w_2 и w_3 ≈ $1{,}55w_2$ и w_4 ≈ 1,57 w_3 и w_4 ≈ 1,57

евклидова метрика для

w_1 и w_2 ≈ $0{,}17w_1$ и w_3 = 0 w_1 и w_4 ≈ 0,36

w_2 и w_3 ≈ $0{,}17w_2$ и w_4 ≈ 0,38 w_3 и w_4 ≈ 0,36

манхэттенское расстояние для

w_1 и w_2 ≈ $0{,}23w_1$ и w_3 = 0 w_1 и w_4 ≈ 0,77

w_2 и w_3 ≈ $0{,}23w_2$ и w_4 ≈ 0,84 w_3 и w_4 ≈ 0,77

Виды алгоритмов кластеризации

- Иерархические и плоские алгоритмы
  - **иерархические** строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений
  - **плоские** строят одно разбиение объектов на кластеры
- Четкие и нечеткие алгоритмы
  - **четкие** каждому объекту выборки ставят в соответствие номер кластера
  - **нечеткие** каждому объекту ставят в соответствие набор вещественных значений – степень отношения объекта к кластерам

Иерархические алгоритмы

Восходящие (агломеративные): построение кластеров снизу вверх

Начало: один документ – один кластер

Последовательно объединяем пары кластеров

В итоге: один кластер – все документы

Нисходящие (дивизивные): построение кластеров сверху вниз

Начало: все документы – один кластер

Рекурсивно делим кластеры пополам

(с помощью алгоритма плоской кластеризации)

В итоге: один кластер – один документ

«История» объединения/деления кластеров дает их иерархию (бинарное дерево)

Восходящие алгоритмы: критерии объединения

Сходство двух кластеров есть:

- сходство между их наиболее похожими документами (одиночная связь)
  - ✔ создаются протяженные кластеры
  - ✔ не учитывает вся структура кластера
- сходство между их наиболее непохожими документами (полная связь)
  - ✔ создаются компактные кластеры
  - ✔ учитывает вся структура кластера
- среднее сходство всех пар документов (групповое усреднение)
- сходство между их центроидами

Вопросы

Какой тип сходства изображен на Рисунке 1?

Какой тип сходства изображен на Рисунке 2?

Какие должны быть рисунки для других типов?

Рисунок 1

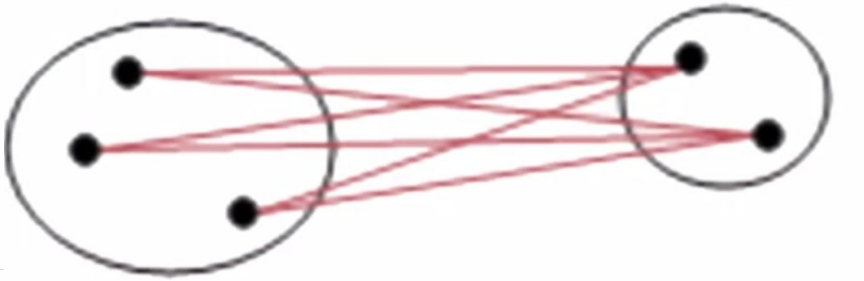


Рисунок 2

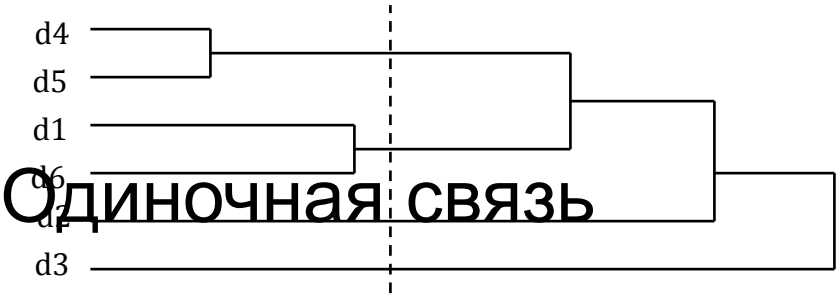
Пример (1): тексты

№	Термины в документы	c=«Китай»
1	китайский пекин китайский	*c*
2	китайский китайский шанхай	*c*
3	китайский макао	*c*
4	токио япония китайский	¬*c*
5	китайский китайский китайский токио япония	?
6	токио пекин	?

Пример (1): деревья

Матрица расстояний:

sim	d_1	d_2	d_3	d_4	d_5	d_6
d_1	0					
d_2	1,36	0				
d_3	1,37	1,39	0			
d_4	1,36	1,39	1,40	0		
d_5	1,32	1,36	1,38	0,27	0	
d_6	0,66	1,43	1,41	1,30	1,21	0



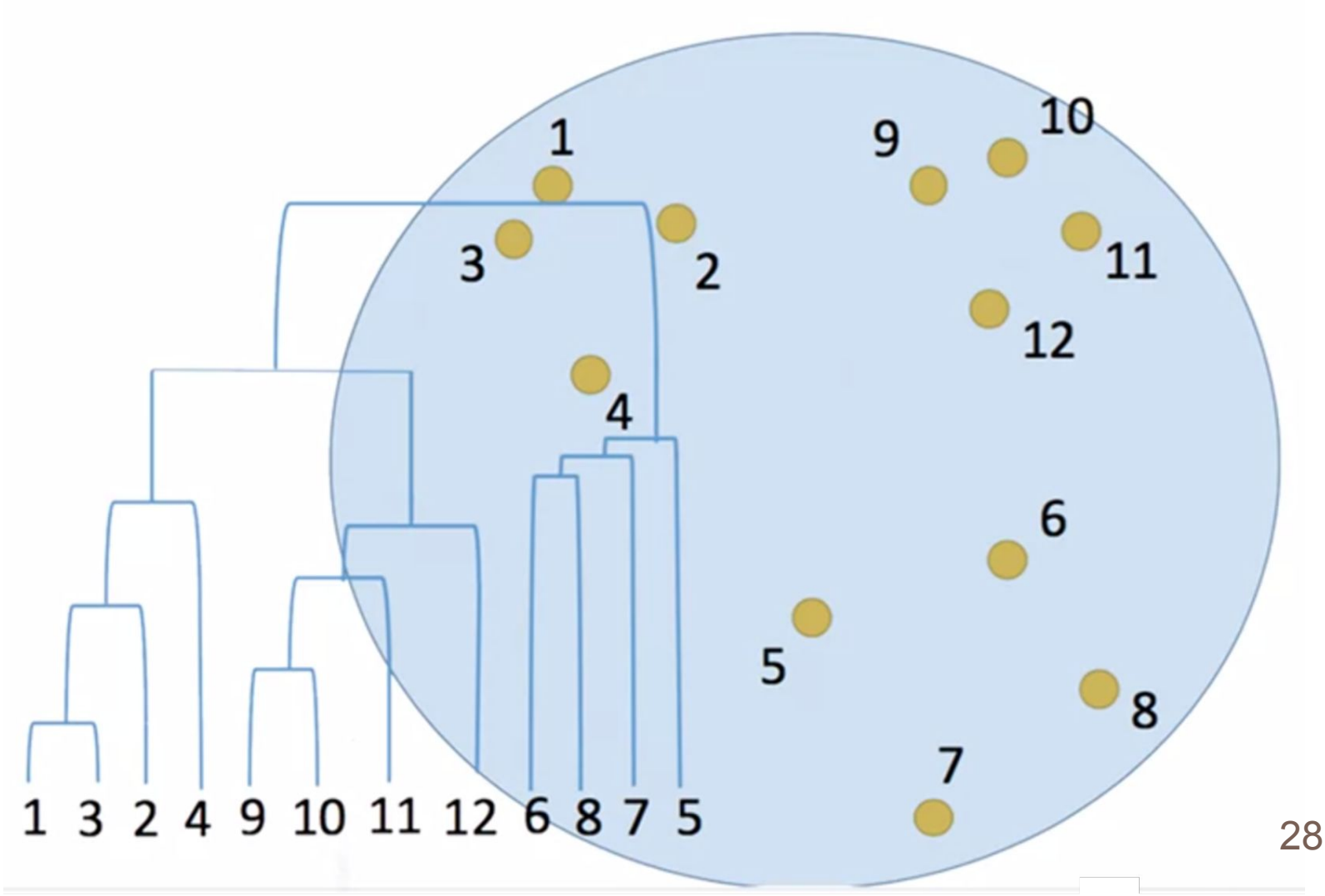
Одиночная связь

Полная связь

Пример 2

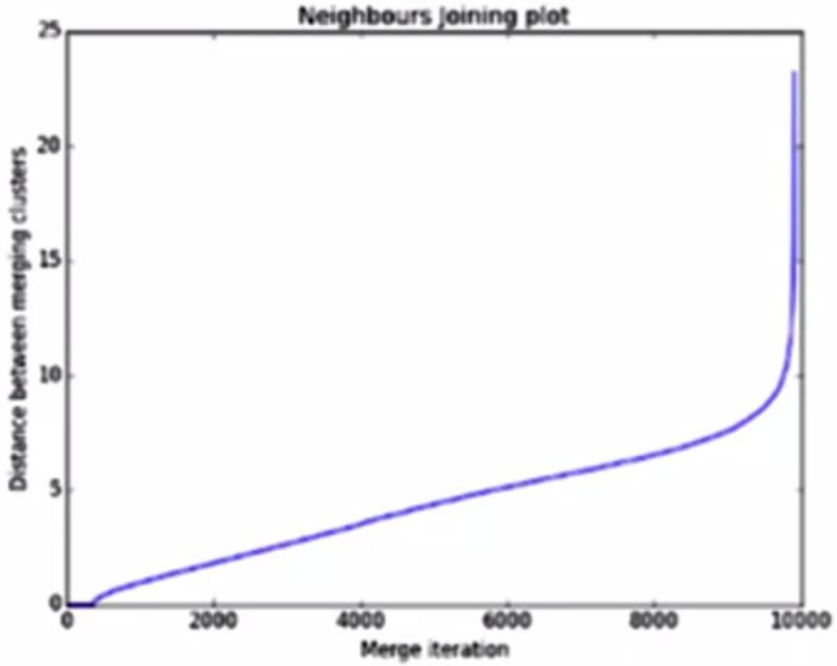
Давайте построим дерево

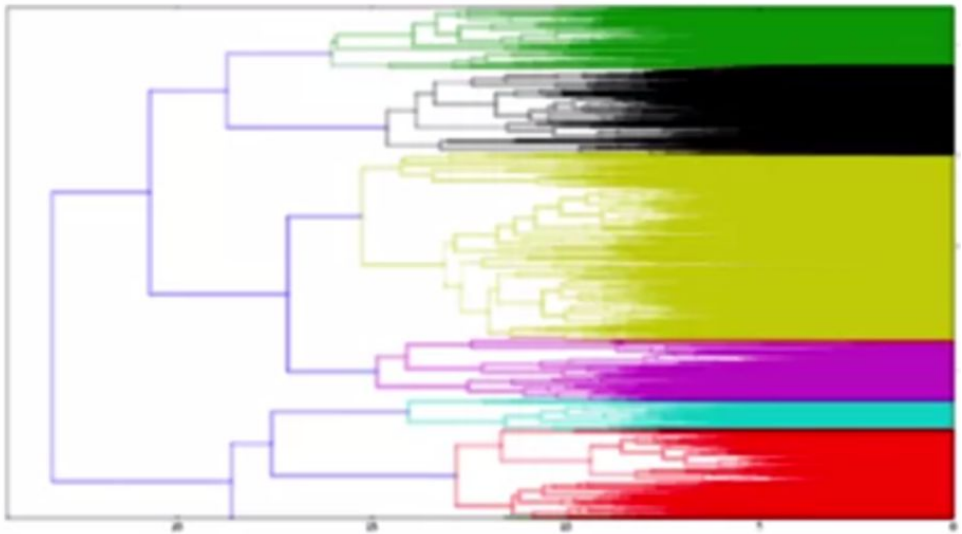
Пример 2: полученное дерево



Просто пример дерева

Выборка из 10000 писем: дерево (дендрограмма) и график зависимости расстояния между объединяемыми кластерами от номера итерации





Плоский четкий алгоритм k-средних (k-means)

Входные данные:

- количество кластеров k
- множество документов как векторов $d_i = (d_{i1}, \ldots, d_{i|T|})$

Выполнение алгоритма:

1. Выбираем k начальных центроидов кластеров
2. Каждый документ относим к тому кластеру, чей центроид является наиболее близким
3. Выполняем повторное вычисление центроидов каждого кластера

Повторяем, пока не достигнем условия остановки:

- достигнуто пороговое число итераций
- центроиды кластеров больше не изменяются
- достигнуто пороговое значение целевой функции

Оптимизируемая функция

Алгоритм минимизирует среднее внутрикластерное расстояние

- каждая точка присваивается к тому кластеру, центр которого ближе
- каждый центр переходит в среднее арифметическое точек кластера

$$e(D,C)=\sum_{j=1}^{k}\sum_{i:d_i\in c_j}\left\|d_i-\mu_j\right\|^2$$

где μ_j – центроид кластера c_j

$$\mu_j=\frac{1}{|c_j|}\sum_{i:d_i\in c_j}d_i$$

$|c_j|$ – количество документов в кластере c_j

Иллюстрация работы k-средних, k=2

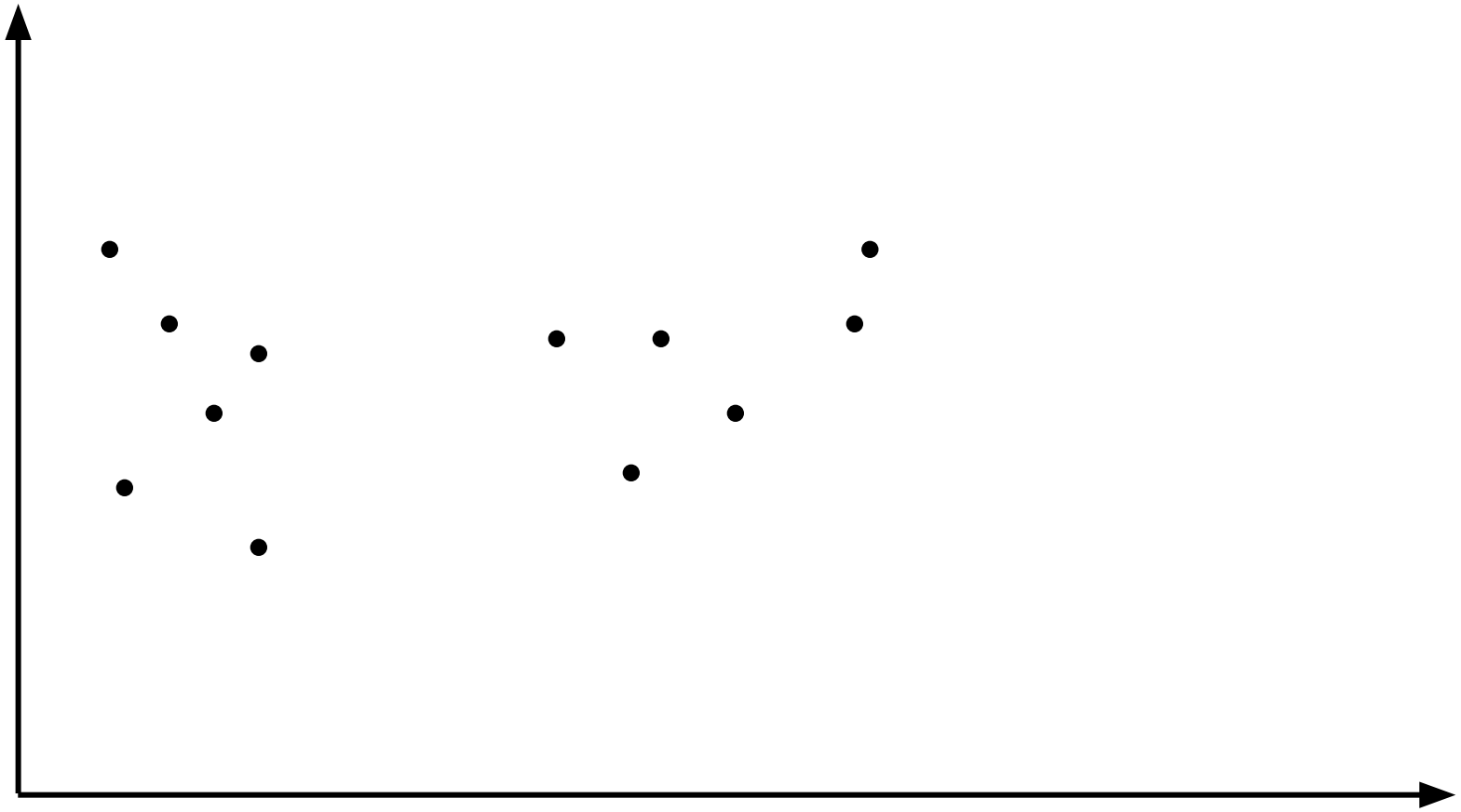


Иллюстрация работы k-средних, k=2

1. Выбираем центроиды (случайным образом)

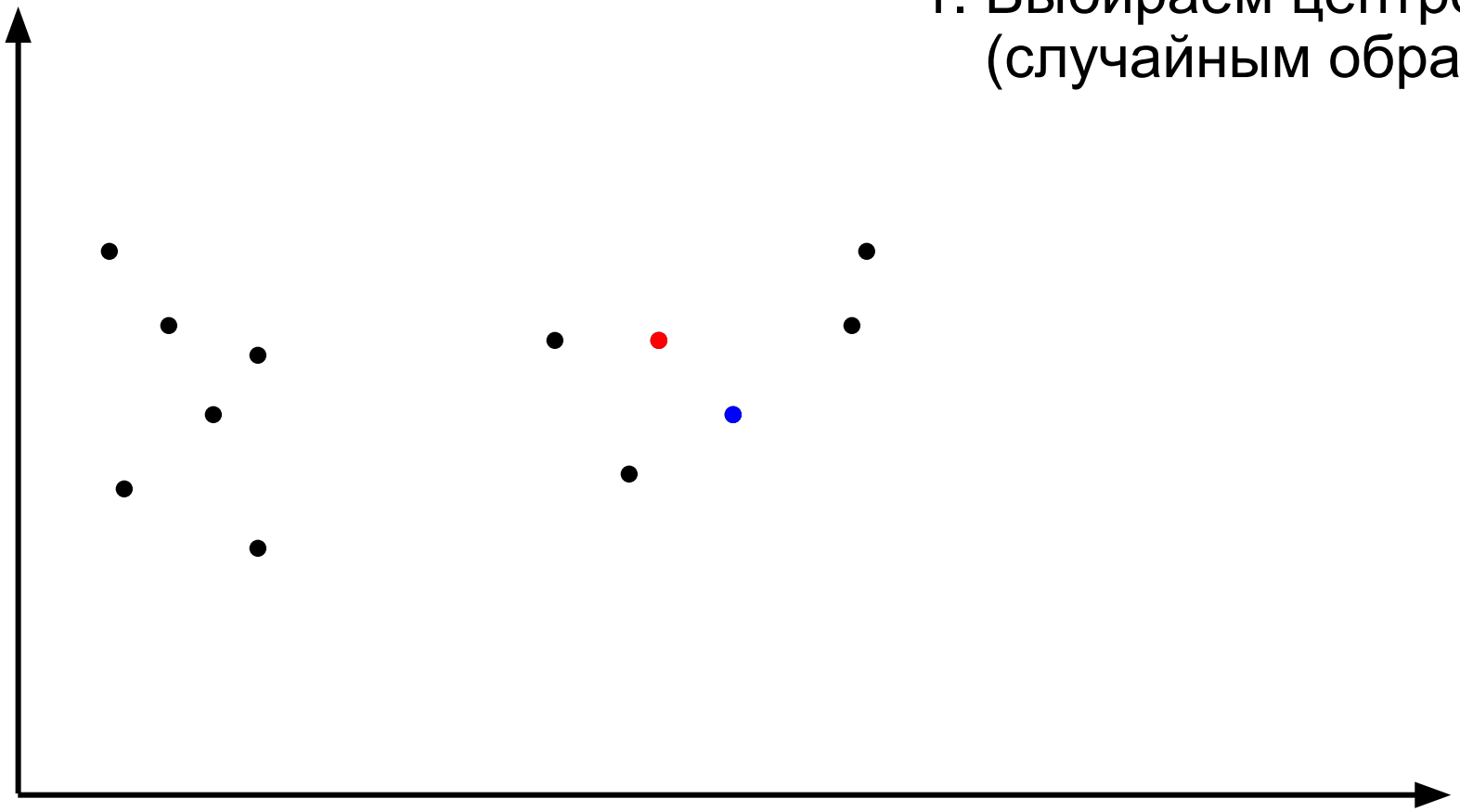


Иллюстрация работы k-средних, k=2

1. Выбираем центроиды (случайным образом)
2. Назначаем кластеры

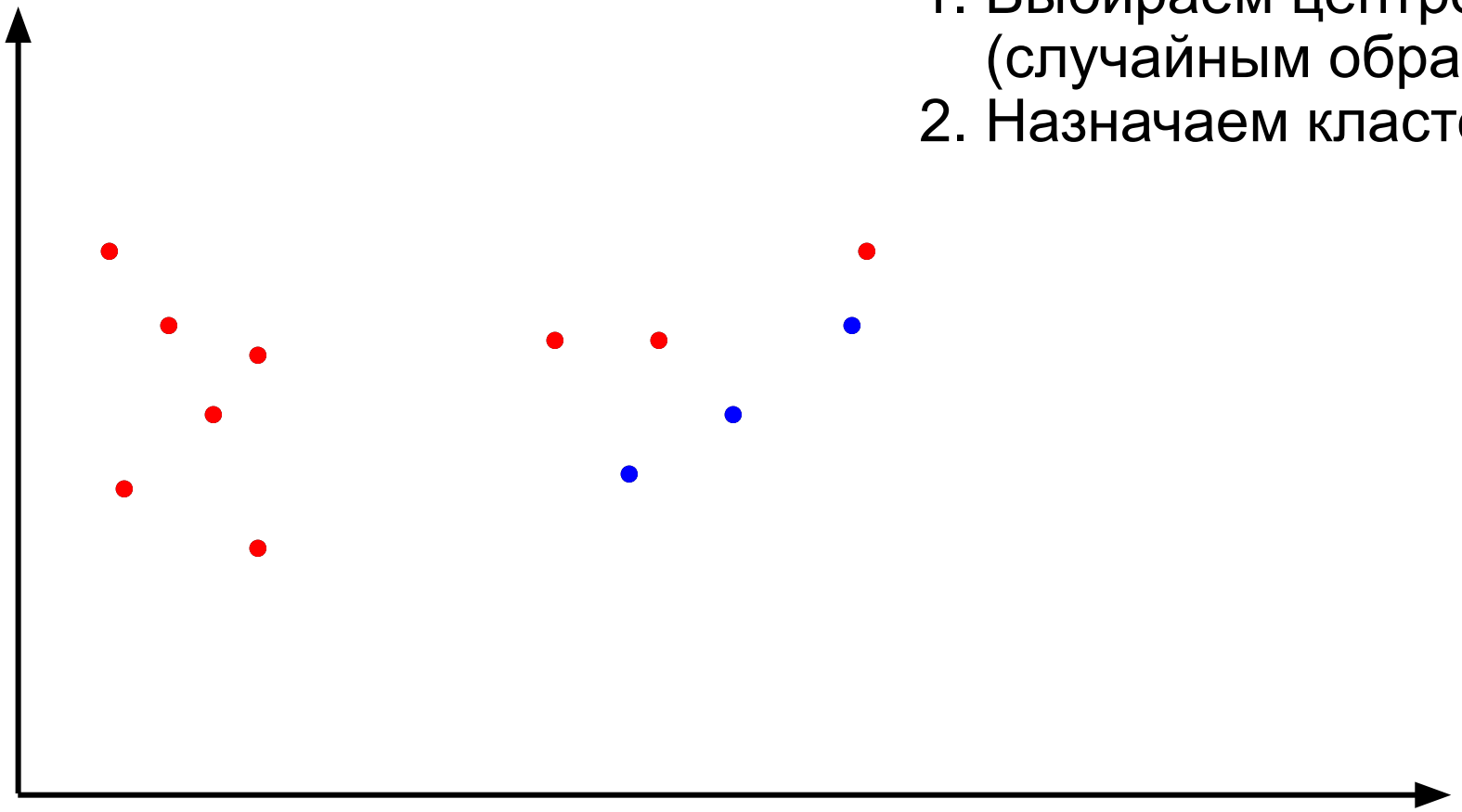


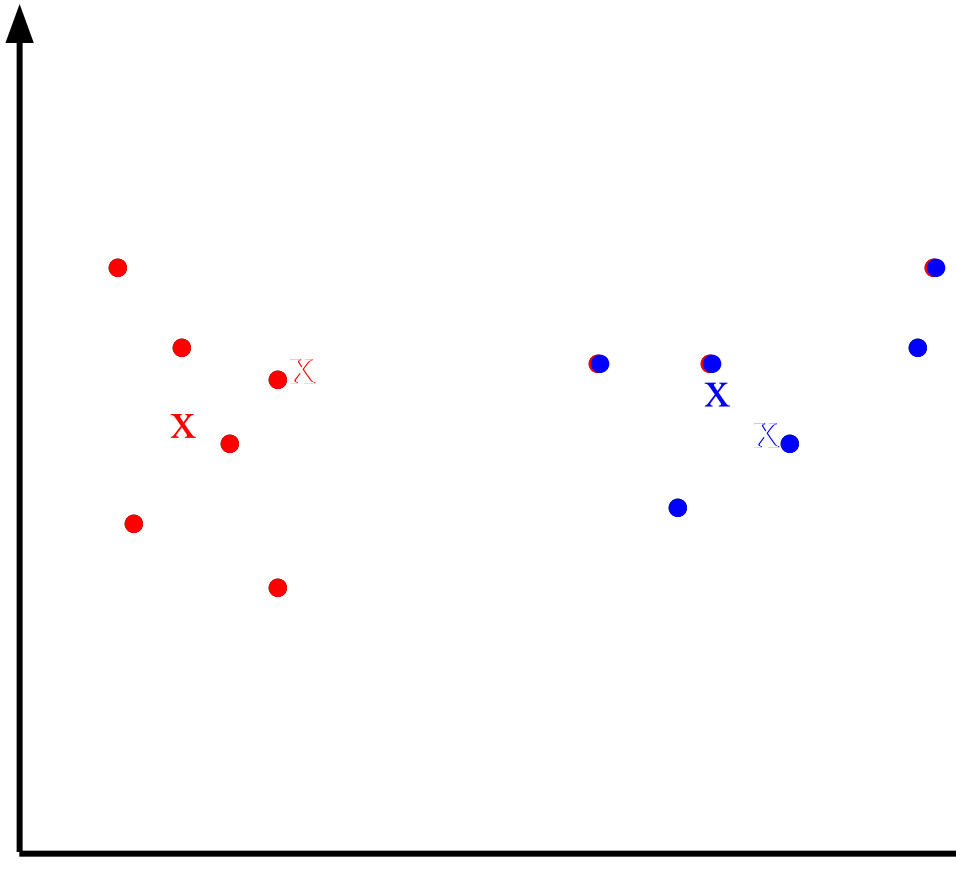
Иллюстрация работы k-средних, k=2

1. Выбираем центроиды (случайным образом)
2. Назначаем кластеры
3. Вычисляем новые центроиды

Иллюстрация работы k-средних, k=2

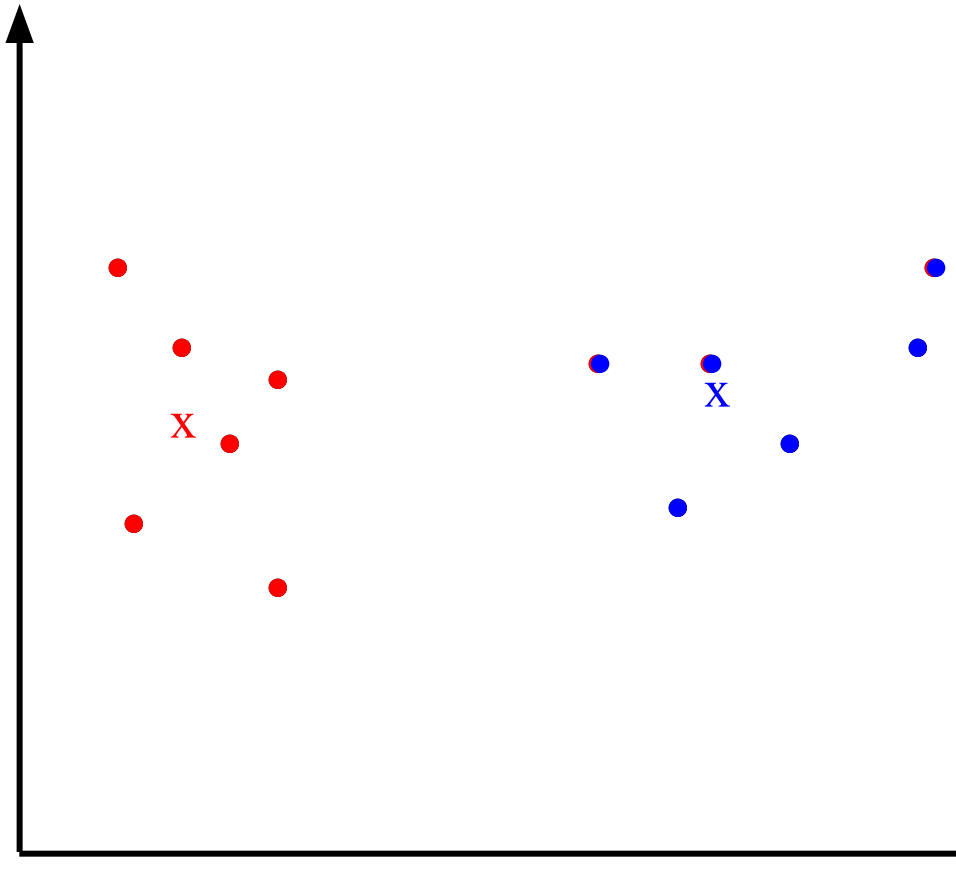
1. Выбираем центроиды (случайным образом)
2. Назначаем кластеры
3. Вычисляем новые центроиды
4. Переназначаем кластеры

Иллюстрация работы k-средних, k=2



1. Выбираем центроиды (случайным образом)
2. Назначаем кластеры
3. Вычисляем новые центроиды
4. Переназначаем кластеры
5. Вычисляем новые центроиды

Иллюстрация работы k-средних, k=2



1. Выбираем центроиды (случайным образом)
2. Назначаем кластеры
3. Вычисляем новые центроиды
4. Переназначаем кластеры
5. Вычисляем новые центроиды
6. Переназначаем кластеры

Иллюстрация работы k-средних, k=2

Пример использования: документы

№	Термины в документы	c=«Китай»
1	китайский пекин китайский	*c*
2	китайский китайский шанхай	*c*
3	китайский макао	*c*
4	токио япония китайский	¬*c*
5	китайский китайский китайский токио япония	?
6	токио пекин	?

Пример использования: применение алгоритма

Итерация 1. Случайным образом инициализированы μ_j:

μ_1=[0,96 0,80 0,42 0,79 0,66 0,85] μ_2=[0,49 0,14 0,91 0,96 0,04 0,9

dist	d_1	d_2	d_3	d_4	d_5	d_6
μ_1	**1,55**	1,81	1,66	**1,51**	**1,38**	**0,85**
μ_2	1,82	**1,38**	**1,37**	1,74	1,59	0,93

- c_1:={d_1, d_4, d_5, d_6} c_2:={d_2, d_3}

dist	d_1	d_2	d_3	d_4	d_5	d_6
μ_1	**0,74**	1,21	1,22	**0,68**	**0,61**	**0,67**
μ_2	1,18	**0,69**	**0,69**	1,21	1,18	1,24

И

[.9 0 0]

Пример использования: уменьшение цветов изображения

Охарактеризуйте рисунки с точки зрения цвета

Пример использования: уменьшение цветов изображения

96615 цветов

чайно)

64 цвета (K-means)

Проблемы алгоритма k-средних

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *e(D,C)*
- Результат зависит от выбора исходных центров кластеров. Решение – эвристические правила (например, алгоритм k-means++)
- Число кластеров надо знать заранее. Решение – специальные методы (например, метод локтя)
- Не справляется с задачей, когда объект принадлежит к разным кластерам в равной степени или не принадлежит ни одному. Решение – нечеткие алгоритмы (например, c-means)
- Хорошо работает только для кластеров, близких к сферическим. Решение – другие алгоритмы (например, DBSCAN)

Плоский нечеткий алгоритм c-средних (c-means)

Является модификацией метода k-средних

Входные данные:

- количество кластеров k
- степень нечеткости m
- множество документов как векторов $d_i = (d_{i1}, \ldots, d_{i|T|})$

Выполнение алгоритма:

1. Выбираем k начальных центроидов кластеров
2. Каждый документ относим к тем m кластерам, чьи центроиды являются наиболее близким
3. Выполняем повторное вычисление центроидов каждого кластера
4. Повторяем, пока не достигнем условия остановки

Пример применения: тексты

№	Термины в документы	c=«Китай»
1	китайский пекин китайский	*c*
2	китайский китайский шанхай	*c*
3	китайский макао	*c*
4	токио япония китайский	¬*c*
5	китайский китайский китайский токио япония	?
6	токио пекин	?

	d_1	d_2	d_3	d_4	d_5	d_6
c_1	0,1589	0,6297	0,1369	0,1489	0,0024	0,1920
c_2	0,2860	0,0168	0,0933	0,0843	0,6832	0,3314

Оценка качества кластеризации

Вычисляются меры двух видов:

- Внешние меры: сравнение созданного разбиения с «эталонным»
  - анализируется сходство предсказаний экспертов и предсказаний системы относительно принадлежности каждой пары объектов одному или разным кластерам
- Внутренние меры: анализ внутренних свойств
  - компактность: члены одного кластера должны быть близки друг другу
  - отделимость: кластеры должны далеко отстоять друг от друга

Сравнение алгоритмов кластеризации

Решение задачи кластеризации принципиально неоднозначно:

- не существует однозначно наилучшего критерия качества кластеризации
- количество кластеров заранее неизвестно
- результат кластеризации существенно зависит от того, как определяется схожесть
- нет общепризнанных тестовых данных

Главное основание для выбора алгоритма – знание о его теоретических характеристиках и оценка пригодности для решения поставленной задачи

Домашнее задание. Вариант 1

1. Взять выбранный к прошлому разу набор данных
2. Написать программу кластеризации данных из этого набора. Использовать 2 разных метрики схожести и 2 метода разных видов
3. Подсчитать 2 внутренние и 2 внешние меры качества работы методов
4. Написать отчет, в котором:
   - описать выбранный набор данных
   - описать выбранные метрики, методы кластеризации (не рассказанные – подробно), меры качества
   - сравнить качество работы методов между собой
   - сделать выводы

Домашнее задание. Вариант 2

Написать программу определения расстояния между текстами

1. Взять несколько текстов (около 10)
2. Написать функцию, реализующую одну из мер схожести именно текстов (!!!)
3. Найти попарные расстояния между всеми текстами
4. Написать отчет, в котором:
   - описать структуру программы
   - подробно описать выбранные меры
   - сделать выводы

✔ Стандартные (чужие) меры определения схожести можно использовать только для сравнении их качества работы со своей

Домашнее задание. Вариант 3

1. Найти готовые средства визуализации многомерных векторов (текстов)
2. Рассказать про них на следующем занятии
3. Показать их работу

Спасибо за внимание!