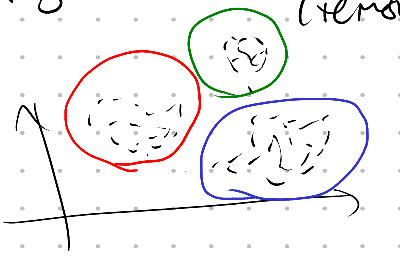
Обучение без учителя.

- набор данных, объекты   f(объект) = значение → дискр / → непр.
- нет объектов, для которых знаем ответы.

Если объекты = тексты, сами разбиваем на классы. (темы)



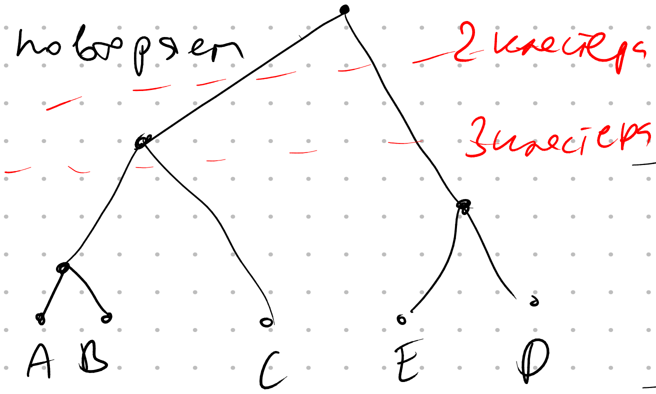
Наглядно для объекта = точки
кластеры - группы соседних точек.

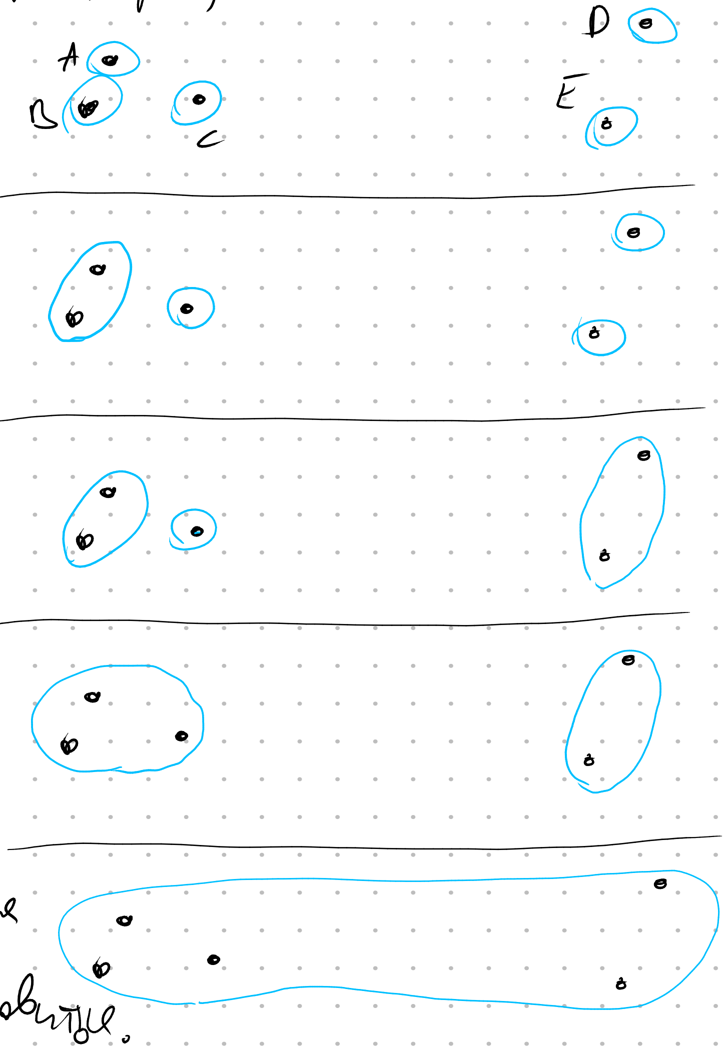
1. Иерархическая кластеризация.

Начинаем с n точек, каждая точка = свой кластер

Находим самые близкие кластеры, объединяем

повторяем

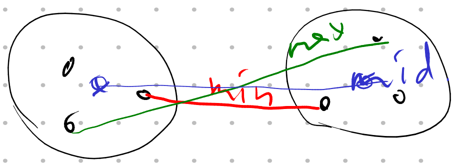




когда остановим процесс, тогда и получили окончательные кластеры

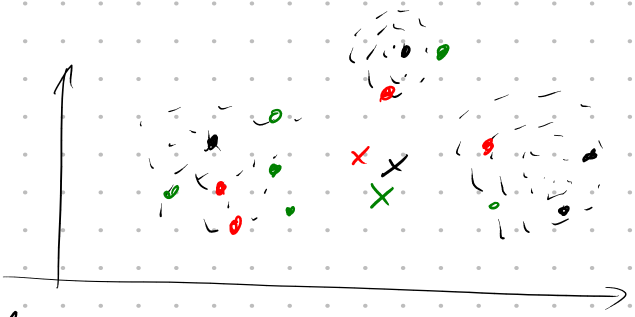
либо заранее знаете, сколько хотите кластеров, либо в процессе работы алгоритма следите, не пора ли остановиться.

Расстояние между кластерами



Другие подходы - знаем, сколько кластеров хотим и сразу разбиваем на столько:

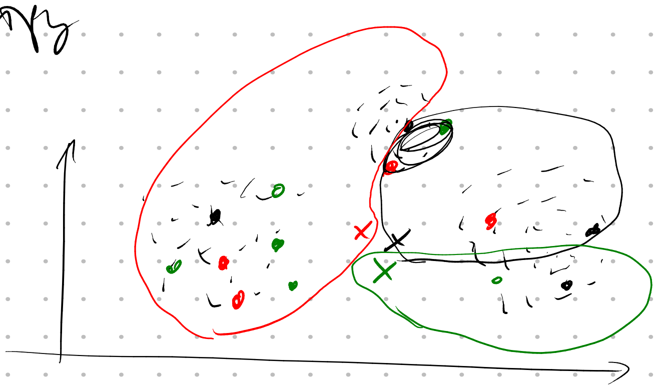
$\boxed{\text{k-средних}}$



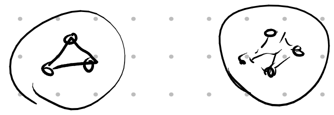
Сначала k случайных кластеров

находим центры, и переназначаем кластеры, находя) точке и ближ. центру

и т. д.



J = качество кластеризации: $\sum$ расстояний внутри кластеров → min



Еще есть спектральная кластеризация.

Но у нас же кластеризуются документы/предложения

↑ для автоматич. реферирования

/слова.

Нужно уметь превращать док/предл/слова в точки (в $\cancel{\mathbb{R}^2}$ $\mathbb{R}^{256}$ $\mathbb{R}^{100000}$).

Такие отображения будем называть: векторизация, embedding

Для документов в архиве:

каждое значение в векторе - это одно слово.

сл1 сл2. . . . . . слN - все слова словаря.

text1

text2

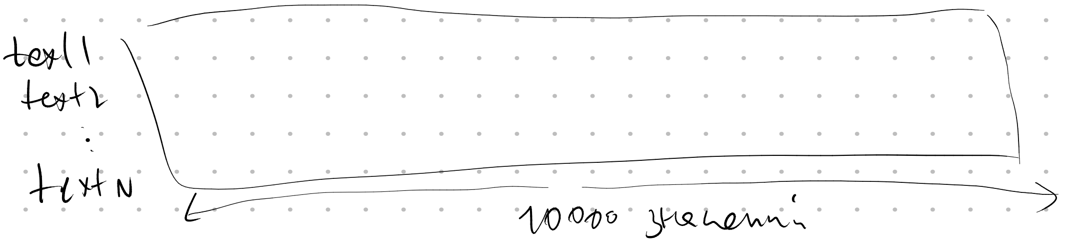
1) частота 10 5 0 0 0 0 1 - абсолютн.

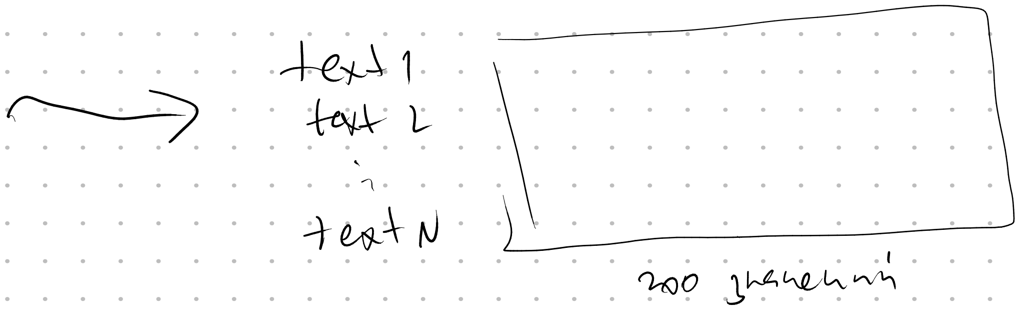
2) частота 0.01 0.001 0 0 0 0 0.0002 - относит.

3) tf-idf $= \underbrace{\text{частота слова}}_{\text{term frequency}} \cdot \underbrace{\text{обр частот по док}}_{\frac{1}{\log(\text{doc freq}+1)}}$

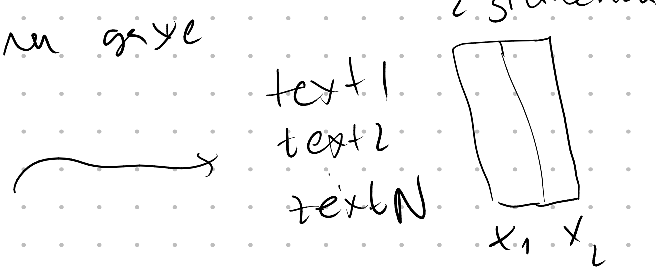
. . .

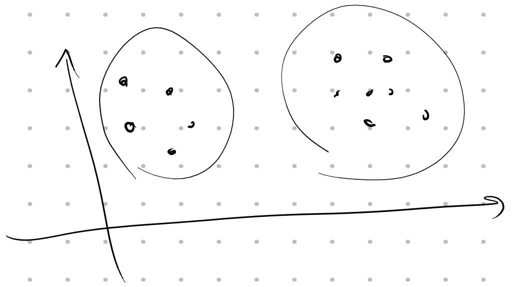
есть методы уменьшения размерности





или даже

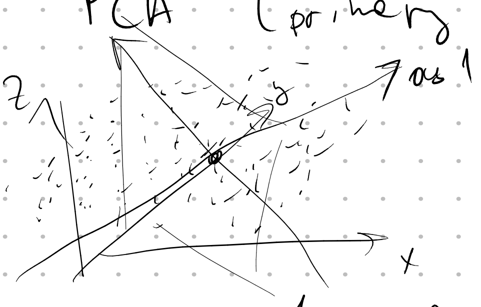


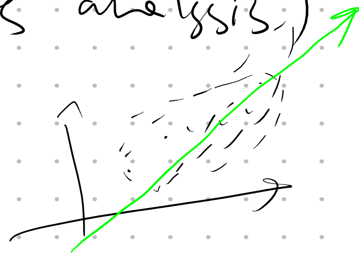


Примеры задач.

1. Кластеризация документов. новости → темы
tf-idf

+ нарисовать их на плоскости.
привести к размерности 2 методом
PCA (primary components analysis)

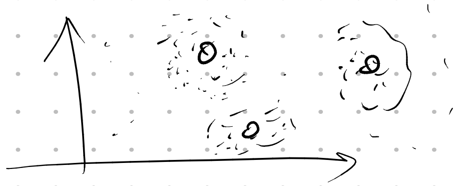




кластеризовать можно
1. в исх пространстве
✓ 2. в пр-ве после PCA →100
3. — // — →2

2. Кластеризация предложений из текстов двух авторов. Совпадёт ли результат кластеризации с реальным разбиением на авторов?

3. Документ. Кластеризуем предложения.
Реферат документа — это набор предложений по одному из кластера.



4. Смыслы слов.
Толковый словарь

коса
— длинные волосы
— инструмент
— ...

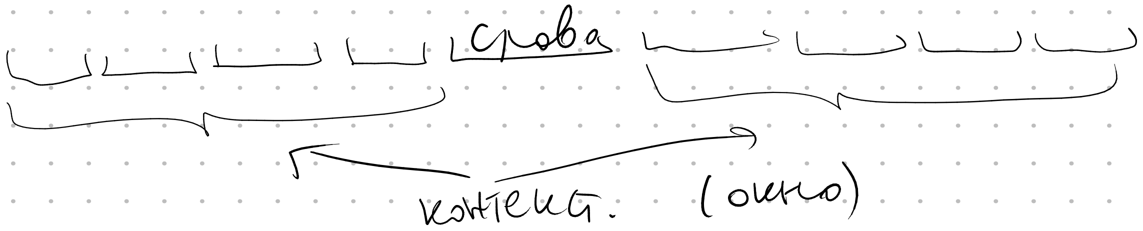
идти - перемещается ногами
- происходит
- красиво. сочетается
-
- :

то |

можно определять возможные смыслы автоматически.

Корпус, где используется слово:

применение слова



все применения одного слова можно кластеризовать по контекстам. кластер = смысл.

как векторизовать контекст ?

1) tf-idf.

2) word2vec (позже)

какие слова использовать?

1) "идти" - многозначное слово с тонкими различиями смыслов.

2) "коса" "замок" явные разные смыслы

3) "банан" + "шар" ← как будто это одно слово

до 11:20