Topic modelling and dimensionality reduction

masha.shejanova@gmail.com

Тематическое моделирование

Что и зачем

Тема — “о чём документ” ≃ набор часто совместно встречающихся слов

Мы считаем, что тема употребление того или иного слова зависит от темы. А тема — от документа.

Зачем:

- поиск в электронных библиотеках
- трекинг новостных сюжетов
- “продвинутый” эмбеддинг документа

topic modeling vs. clustering

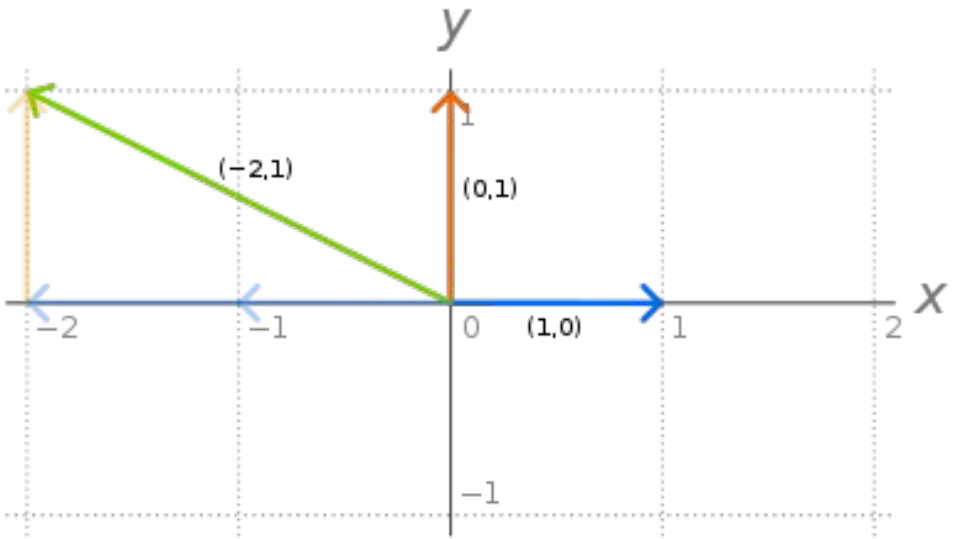
Что похожего: есть документы, раскидываем их по кучкам, заранее не знаем по каким.

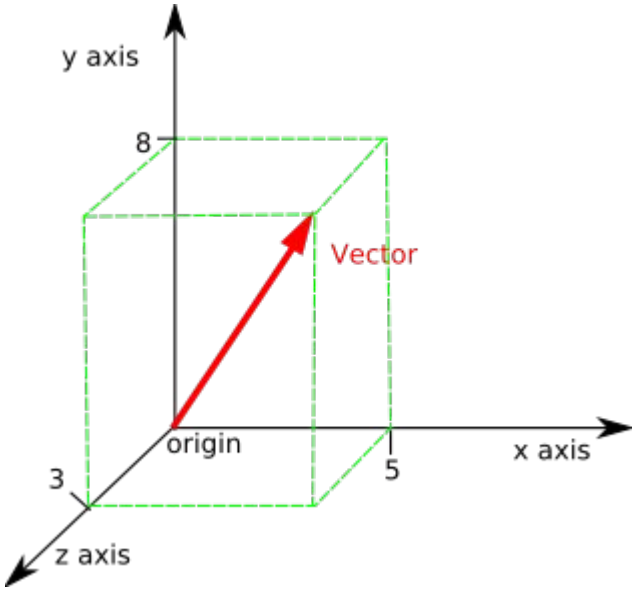
Что разного: у одного документа может быть высокая степень принадлежности больше, чем к одной теме.

PCA (метод главных компонент): идея

Базис линейного пространства

Стандартный базис:





Замена базиса

На самом деле, базисные вектора можно выбирать как угодно — главное чтобы можно было выразить через них все вектора пространства.

(И чтобы сами базисные вектора нельзя было выразить друг через друга).

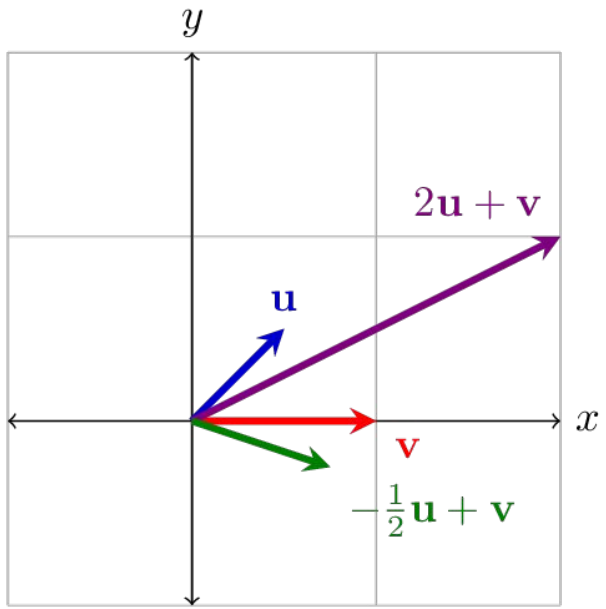
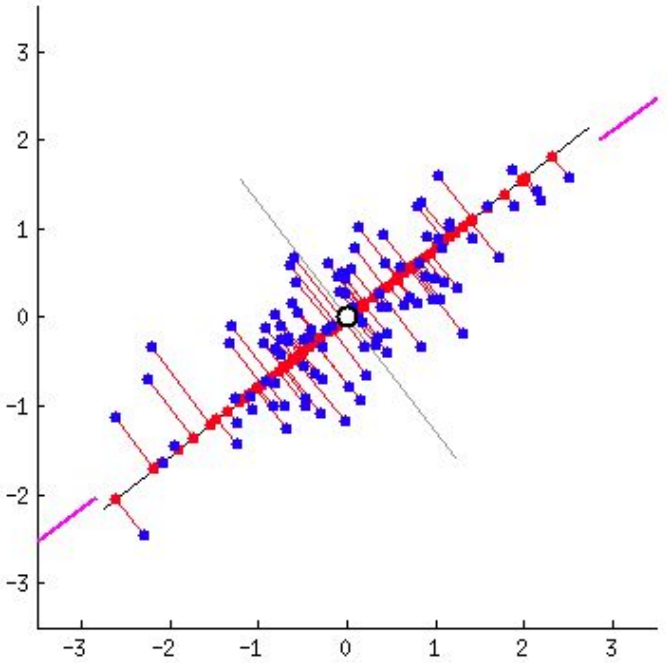


Figure 1: Vector combinations.

PCA

Найдём такой базис, чтобы как можно лучше выразить как можно больше значений за счёт фиксированного количества базисных векторов.

Сделаем проекцию всех данных на эти вектора.



SVD (сингулярное разложение): реализация

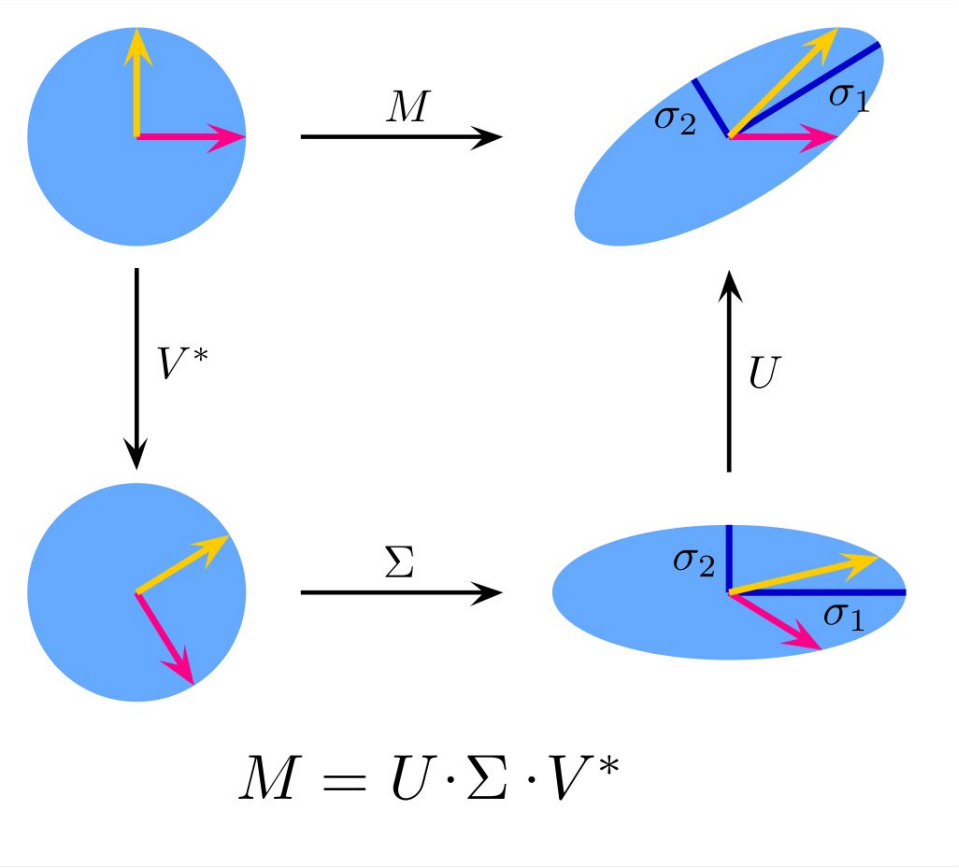
SVD

Любую матрицу M можно разложить на произведение трёх матриц: $M = U * \Sigma * V^*$

U, V^* — матрицы поворота

Σ — матрица растяжения

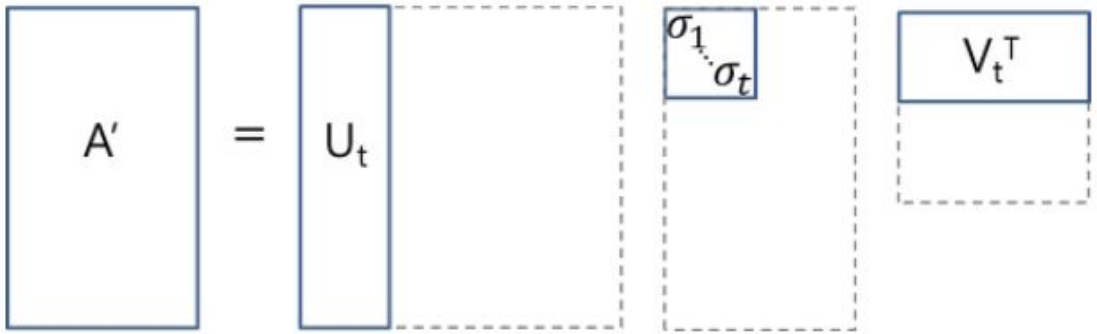
У Σ числа стоят только на главной диагонали, причём они убывают



Truncated SVD

$$A \approx U_t S_t V_t^T$$

Intuitively, think of this as only keeping the t most significant dimensions in our transformed space.

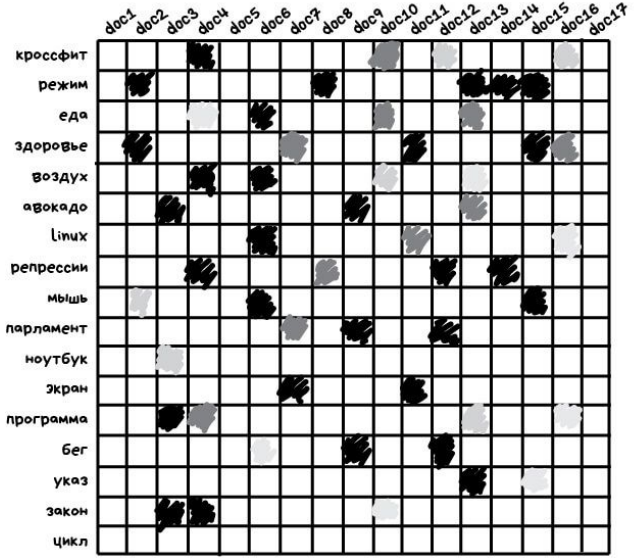


(скрин из вот этой статьи)

Truncated SVD = LSA (latent semantic analysis) in topic modeling

Разделение документов по темам

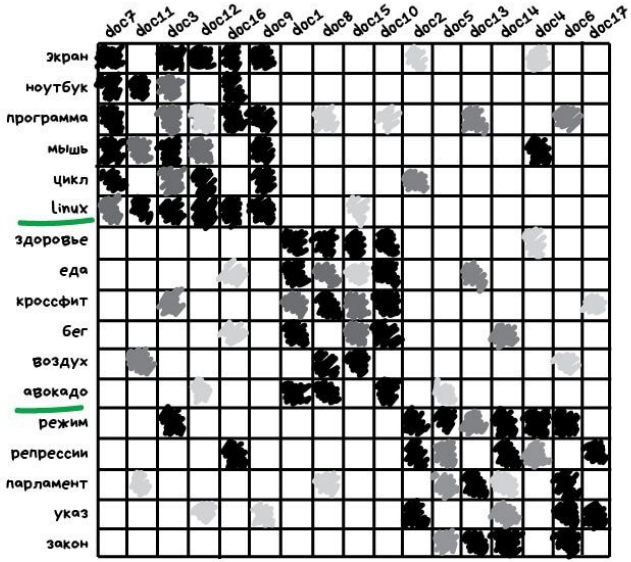
(источник)



1. Строим матрицу как часто каждое слово встречается в каждом документе (чернее - чаще)

SVD

2. Раскладываем



3. Получаем наглядные кластера по тематикам (даже если слова не встречались вместе)

Латентно-семантический Анализ (LSA)

На собачках

При большом количестве компонент разница незаметна.

(источник)

Что ещё бывает? (Более продвинутые вещи)

- pLSA: Probabilistic Latent Semantic Analysis
- LDA (a Bayesian version of pLSA)
- ARTM — LDA, но с регуляризацией
- bigARTM — ARTM с наворотами :) (но вообще, это библиотека, в которой есть все эти методы и больше!)

Снижение размерности

Что и зачем

В общем случае — у нас есть признаковое пространство на много-много измерений (например, мешок слов по корпусу, и каждое слово — признак). Мы хотим “сжать” их как-то так, чтобы потерять минимум информации.

Каждое новое “измерение” — элемент вектора — будут заключать в себе обобщённое представление нескольких элементов из большого вектора.

- убрать несущественные признаки
- тематическое моделирование
- визуализация

SVD

LSA == PCA == Truncated SVD

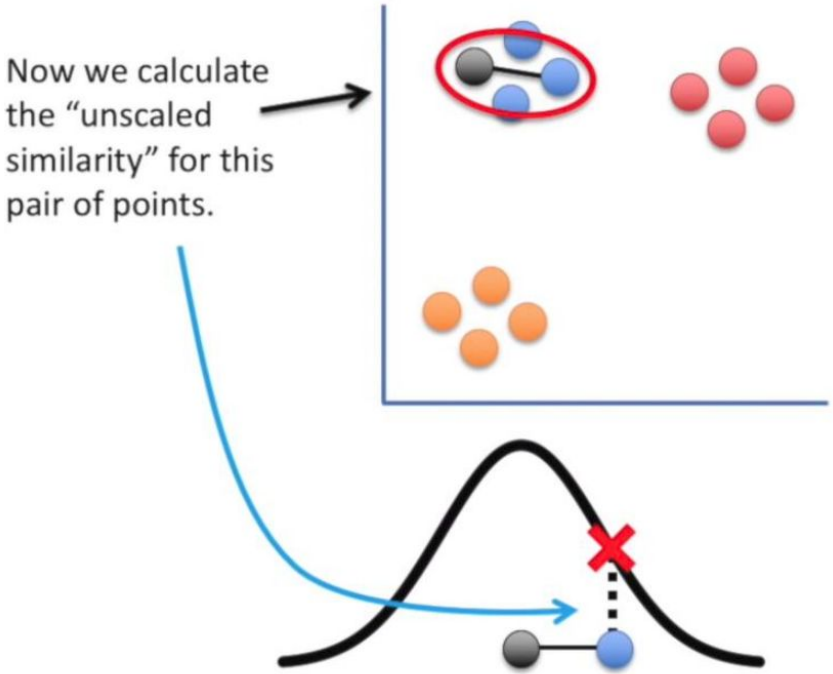
t-SNE

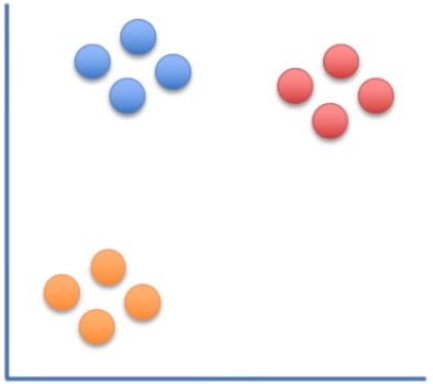
- используется для визуализации
- хорош только для перевода в очень маленькие размерности

Шаги:

- посчитать расстояние от каждой точки до каждой другой (используя формулу нормального (SNE) или T распределения (t-SNE))
- случайно породить соответствующие им точки в маленькой размерности
- решить задачу оптимизации: надо, чтобы распределения расстояний (реальных и в пространстве маленькой размерности) максимально совпадали

t-SNE





At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from…