Дистрибутивная семантика

Квантитативный анализ текста

Кирилл Александрович Маслинский

18.04.2022 / 07

ЕУСПб

Дистрибутивная гипотеза

co-occur | collocates

Дистрибутивная гипотеза: Firth

Firth 1935

the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously

Firth 1957

You shall know a word by the company it keeps

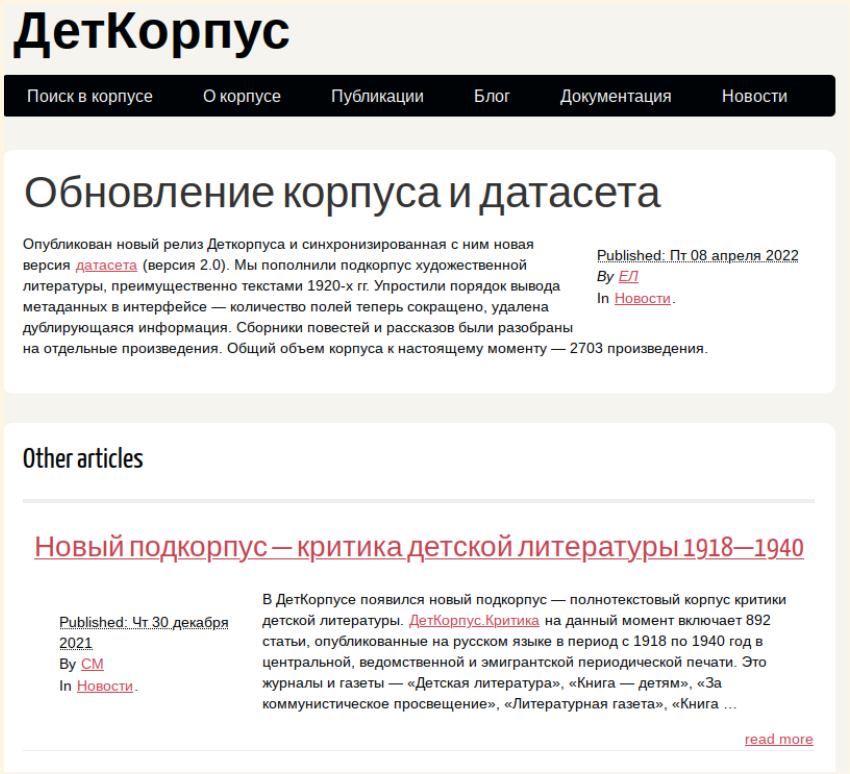
Дистрибутивная гипотеза Зеллига Харриса

Harris 1954

The fact that, for example, not every adjectives occurs with every noun can be used as a measure of meaning difference. For it is not merely that different members of the one class have different selections of members of the other class with which they are actually found. More than that: if we consider words or morphemes A and B to be more different than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference in distribution.

Иллюстрация: дистрибутивная гипотеза в действии

Корпус дня: ДетКорпус



- Корпус прозы для детей и юношества на русском языке
- Период: 1900—2019
- 2573 произведений, 73 млн слов
- Поисковый интерфейс: http://detcorpus.ru

Значение и сочетаемость

	Lemma	Lemma ↓	Частотность	Употреблений на миллион
1	маленький	девочка	668	7,40
2	маленький	ребенок	541	6,00
3	маленький	мальчик	342	3,79
4	маленький	пионер	12	0,13
5	маленький	пионерка	4	0,04
6	долговязый	мальчик	240	2,66
7	долговязый	девочка	3	0,03
8	хороший	мальчик	236	2,62
9	хороший	девочка	234	2,59
10	хороший	ребенок	60	0,67
11	хороший	пионер	18	0,20
12	хороший	пионерка	9	0,10
13	красивый	девочка	136	1,51
14	красивый	мальчик	32	0,35
15	красивый	ребенок	9	0,10
16	рыжий	девочка	129	1,43
17	рыжий	мальчик	42	0,47
18	рыжий	пионер	2	0,02

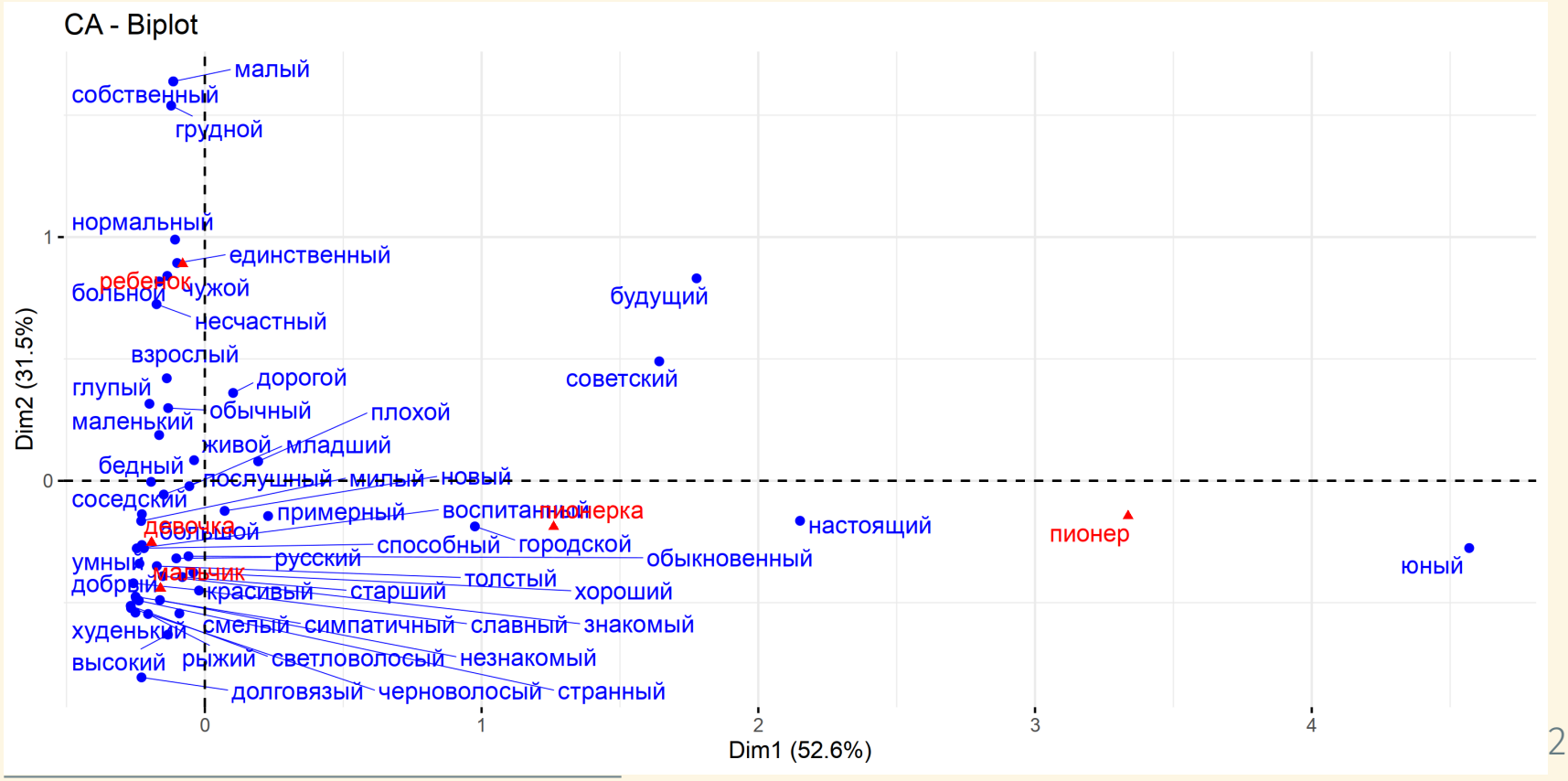
[1]

[1] Статистика из ДетКорпуса: http://detcorpus.ru

Контекстные вектора

	воспитанный	худенький	чужой	долговязый	младший	ю
девочка	85	56	38	3	25	
мальчик	24	40	21	240	52	
пионер	0	0	1	0	3	
пионерка	0	0	0	0	1	
ребенок	15	3	107	0	32	

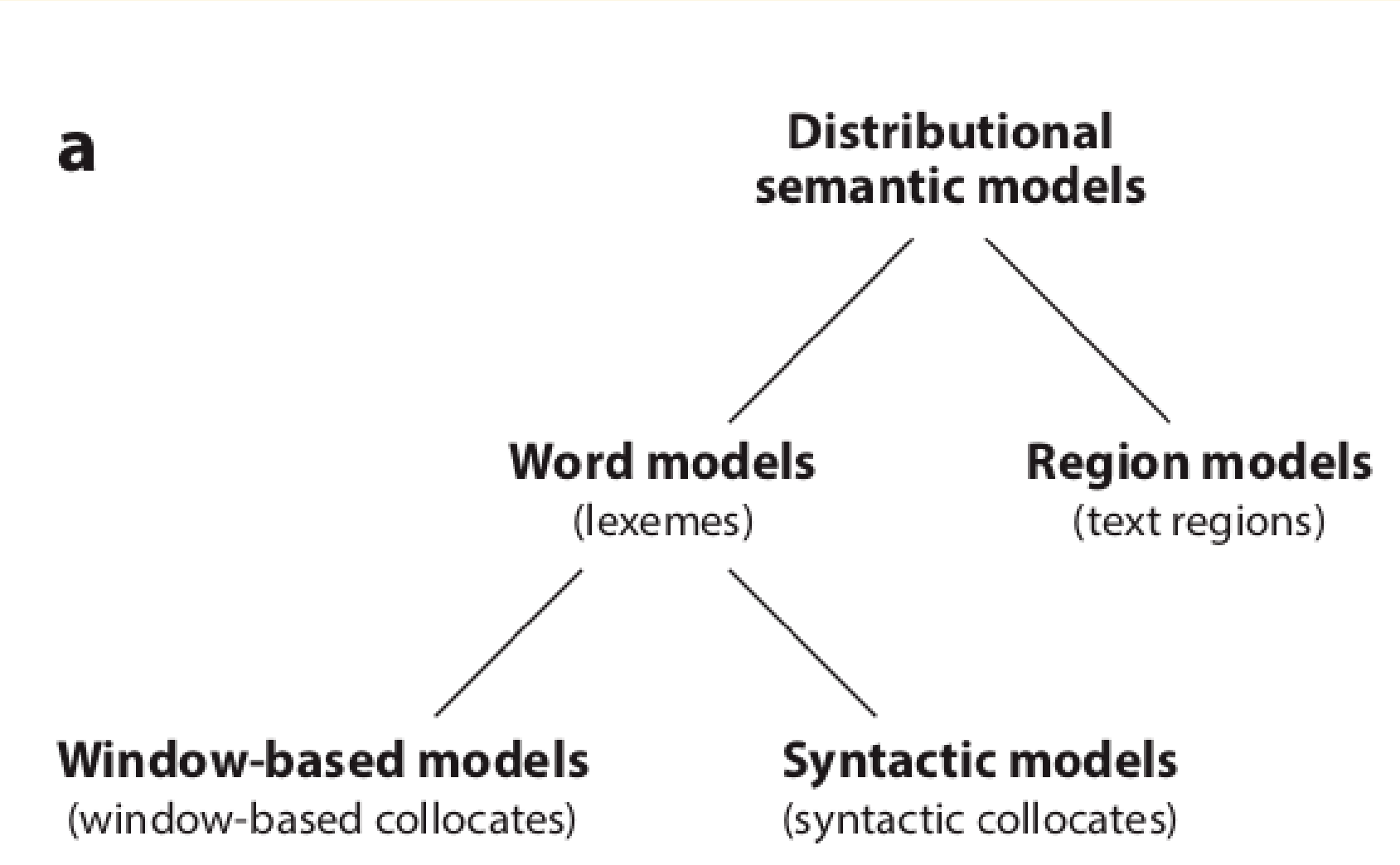
Дистрибуция в пространстве



[2]Анализ соответствий (Correspondence Analysis), выполненный по данным о распределении 50 самых частотных прилагательных в контексте слов ребенок/мальчик/девочка/пионер/пионерка, по данным ДетКорпуса

Разновидности дистрибутивного анализа

Дистрибутивные данные



Контекстное окно

Левый контест	KWIC	Правый контекст
? — — Моя сестра . Когда я улетала на Землю , она была еще	**девочкой**	. Самой младшей во всей нашей колонии . Я ее очень любила
. Девочка сидела на скамейке , держала на коленях книгу , а	**мальчики**	, оба в коричневых сандалиях и матросских рубашках с сини
н снимал девочку в бантиках с пожарной лестницы и как эта	**девочка**	потом решила , что он её спас , и сказала об этом отцу на « М
ижу сошедшую со стены , обретшую плоть вампиршу – лицо	**девочки**	покрывала страшная бледность , а в глазах появилось злобн
істро . Впрочем , не всегда , — созналась старушка , глядя на	**детей**	ясными глазами . Повернувшись , она взглянула на стоящий у
азала мать . – – При чем здесь ребенок ? – ответил отец . –	**Ребенок**	здесь ни при чем . – – Вот-вот , – подхватила мама , – я ни п
От дядьки ... В это время мимо прошел рыжий веснушчатый	**мальчик**	. Он сделал вид , что споткнулся , и взмахнул рукой к Лешкин
на долю которой выпал тяжкий труд , нищета , рождение 14	**детей**	и кулаки пьяницы мужа . Тихая , набожная , она тем не менее
ту , как она поднимала шум на всё училище . Именно то , что	**девочек**	была всего одна группа , давало Тане неоценимое преимущес
Через несколько дней , после утомительной работы , Люсе с	**девочками**	пришлось сесть за стол с табличной « Для лентяев ». Дежурн
Магнитке . Жена его , дочь Ивана Никитича , поехала с ним .	**Мальчика**	взять с собой на новое , необжитое место не решились , остав
гу их маленькие ноги . Правда , случилось однажды , что эти	**девочки**	перестали верить друг другу , перестали ходить вместе . Но э
. И когда на сборе отряда нахлынуло ребят полно и даже не	**пионеров**	, когда городской рассказал о жизни городских пионеров , об
– — Лика , — громким шепотом позвал Ореш - кин . — Лика ...	**Девочка**	повернула голову , но в этот момент Ромку отвлек Димка . И

Контекстное окно

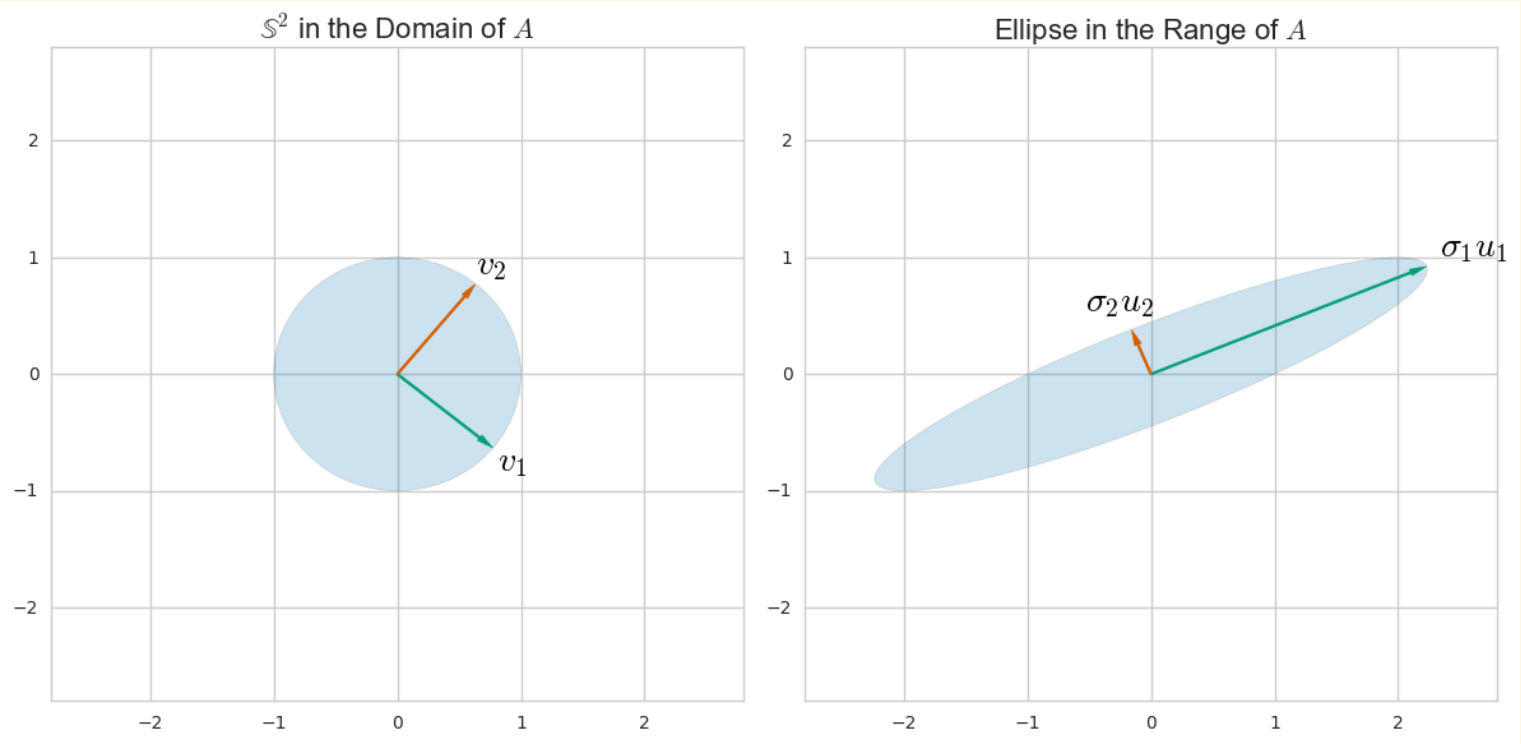
...

вечно просыпал побудку и вылетал на линейку встрепанный и неумытый , о Вадиме , который среди ночи под окнами девочек ухал и хохотал , как настоящий филин , и как девочки своим перепуганным визгом будили весь лагерь , а Вадиму на другой день был общелагерный " влёт "... К Лешке подсел мальчик с тонким , бледным лицом . Лешка заметил его еще утром , потому что мальчик носил очки с толстыми стеклами , и слышал , как ребята называли его " академиком ". — — Меня зовут Яша Брук , — сказал мальчик в очках . — Ты давно у нас ? — — Недавно . — — А откуда ты ? — — Из Батуми . То есть раньше я жил в Ростове , а потом в Батуми . Там и убежал ... — — Убежал ? — — Ага . От дядьки ... В это время мимо прошел рыжий веснушчатый **мальчик** . Он сделал вид , что споткнулся , и взмахнул рукой к Лешкиному лицу . Лешка отпрянул . Рыжий опустил руку , почесал колено , словно он поднимал ее только для этого , и подмигнул ребятам . Те засмеялись , Лешка побледнел и весь напрягся . — — Не приставай , Валет ... — досадливо сказал Яша . — А почему убежал ? Плохо было , да ? Ничего , у нас хорошо , вот увидишь ... Лешка не видел ничего хорошего , он видел теперь только рыжего Валета . Обрадованный своим успехом , тот решил продлить забаву и , снова подмигнув товарищам , направился к Лешке . Однако повторить не удалось . Не успел он взмахнуть рукой , будто бы для того , чтобы почесать затылок , как тут же отдернул ее : Лешка ребром ладони ударил его по плечевому

Методы снижения размерности

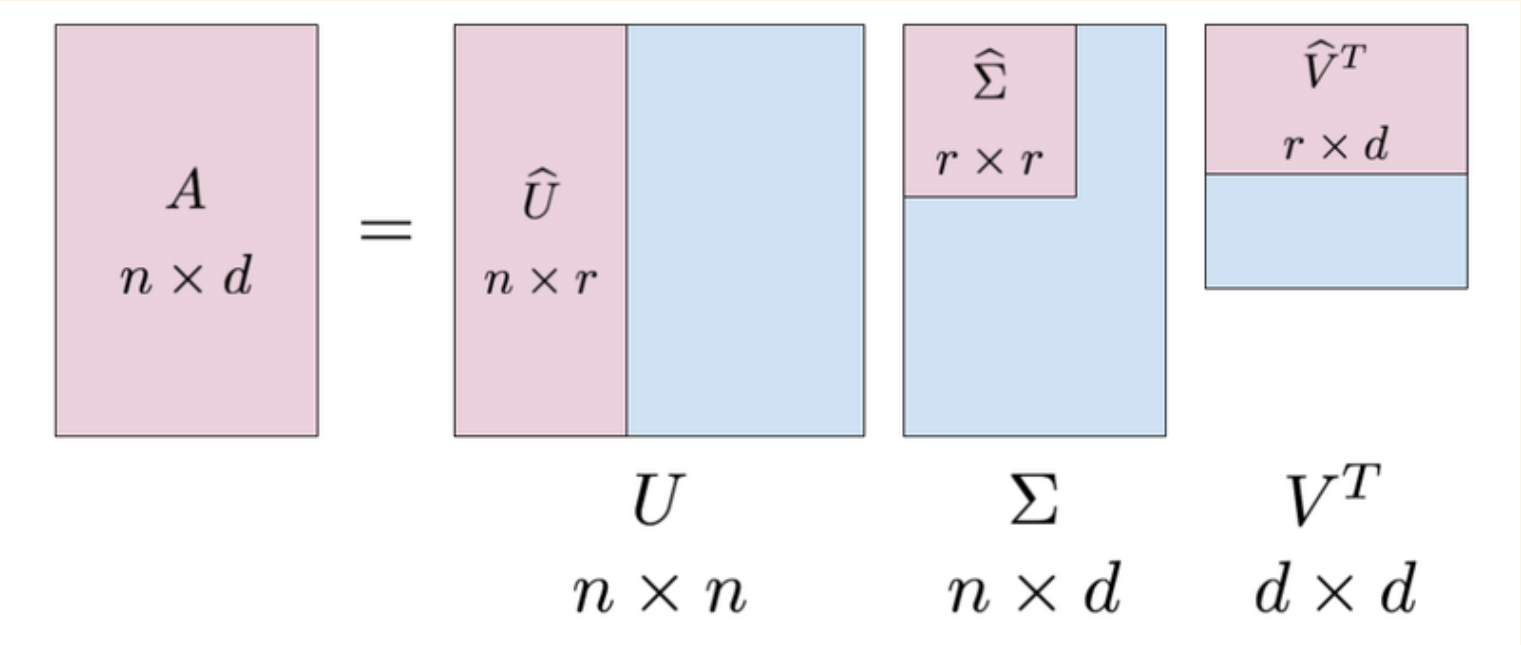
PCA — Principal component analysis

Singular Value Decomposition



- Truncated SVD: производит «сгущенную» матрицу, в которой связанные строки и столбцы частично объединены.
- Измерений (колонок) в результирующей матрице

Singular Value Decomposition



- Truncated SVD: производит «сгущенную» матрицу, в которой связанные строки и столбцы частично объединены.
- Измерений (колонок) в результирующей матрице меньше, чем в исходной.

Латентный семантический анализ (LSA)

LSA

Слова и документы — вектора в семантическом пространстве, измерения которого представляют собой «латентные» переменные.

- совместно встречающиеся слова проецируются на одни и те же измерения;
- вектор для документа — взвешенная сумма векторов входящих в него слов (центроид);
- в семантическом пространстве угол между векторами документов может быть малым, даже если в документах нет общих слов.

Применение LSA

Можно оценивать семантическую близость (сходство):

- слово—слово
- слово—документ
- документ—документ
- документ—слово

Характерные применения LSA:

- поиск документов, близких к запросу (killer feature для информационного поиска);
- выявление семантически близких слов (синонимов);
- оценка связности текста (семантической близости соседних абзацев);
- извлечение ключевых слов текста.