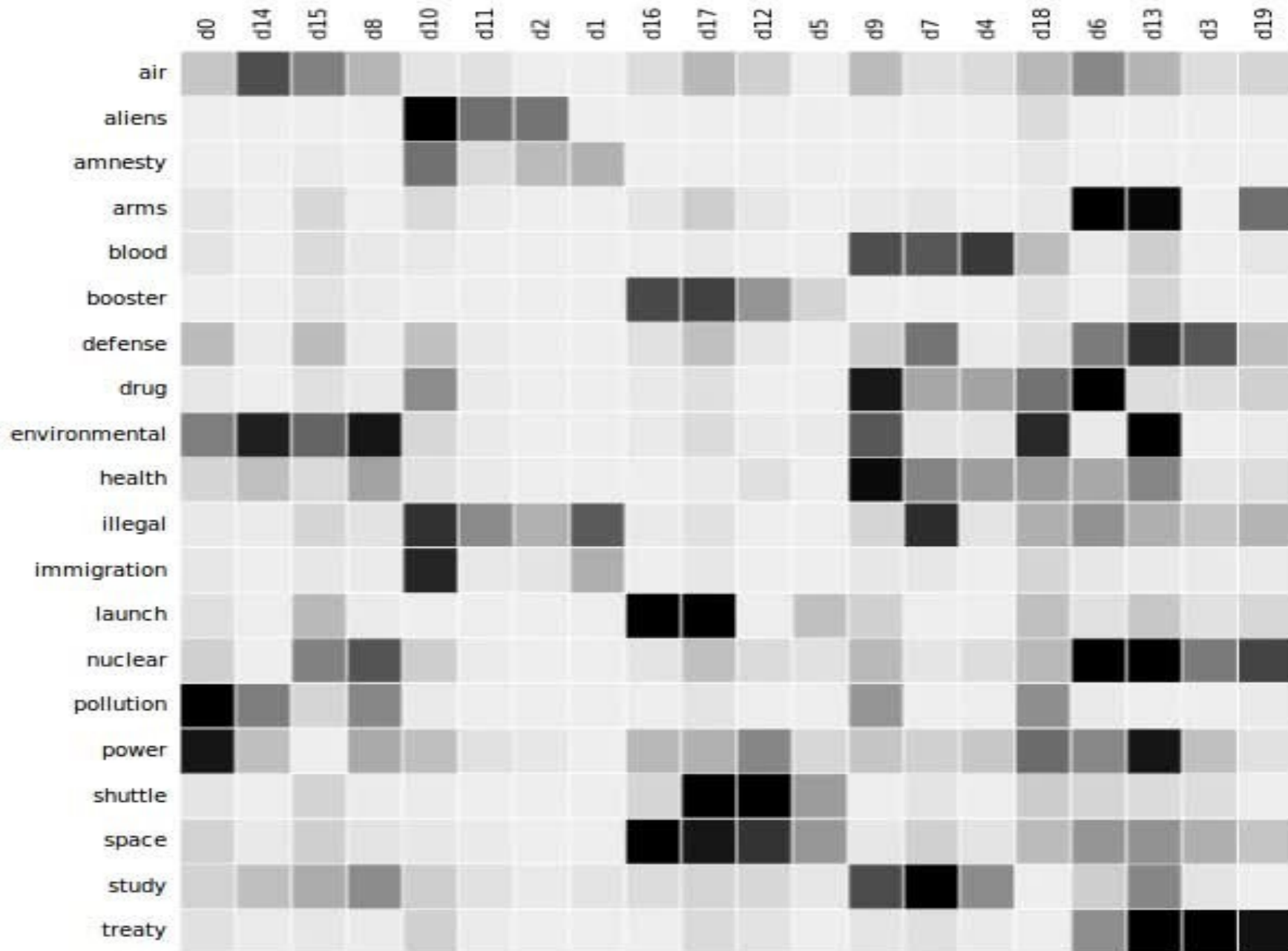


Моделирование семантической близости слов

Векторные модели и др. модели

- Высокая размерность представления документов
- Разреженная матрица терм-документ
- Нет никаких семантических связей между словами
- Сходство между документами определяется только вхождениями одних и тех же слов
- Не учитывается совместная встречаемость слов
- Проблемы со словами:
 - Синонимия – реальное сходство больше, чем по векторной модели
 - Многозначность – реальное сходство меньше, чем по векторной модели



Latent Semantic Analysis (LSA) 1988

- LSI – latent semantic indexing в сфере информационного поиска
- Идея:
 - Перевести термы и документы в пространство меньшей размерности, которое должно отражать семантические связи между словами
 - Сходство между документами высчитывается на основе сходства в пространстве низкой размерности
- Цели:
 - Сходные по смыслу слова должны оказаться в близких точках нового пространства
 - Снижение шума за счет сокращения размерности

LSA: применение сингулярного разложения к матрице терм-документ

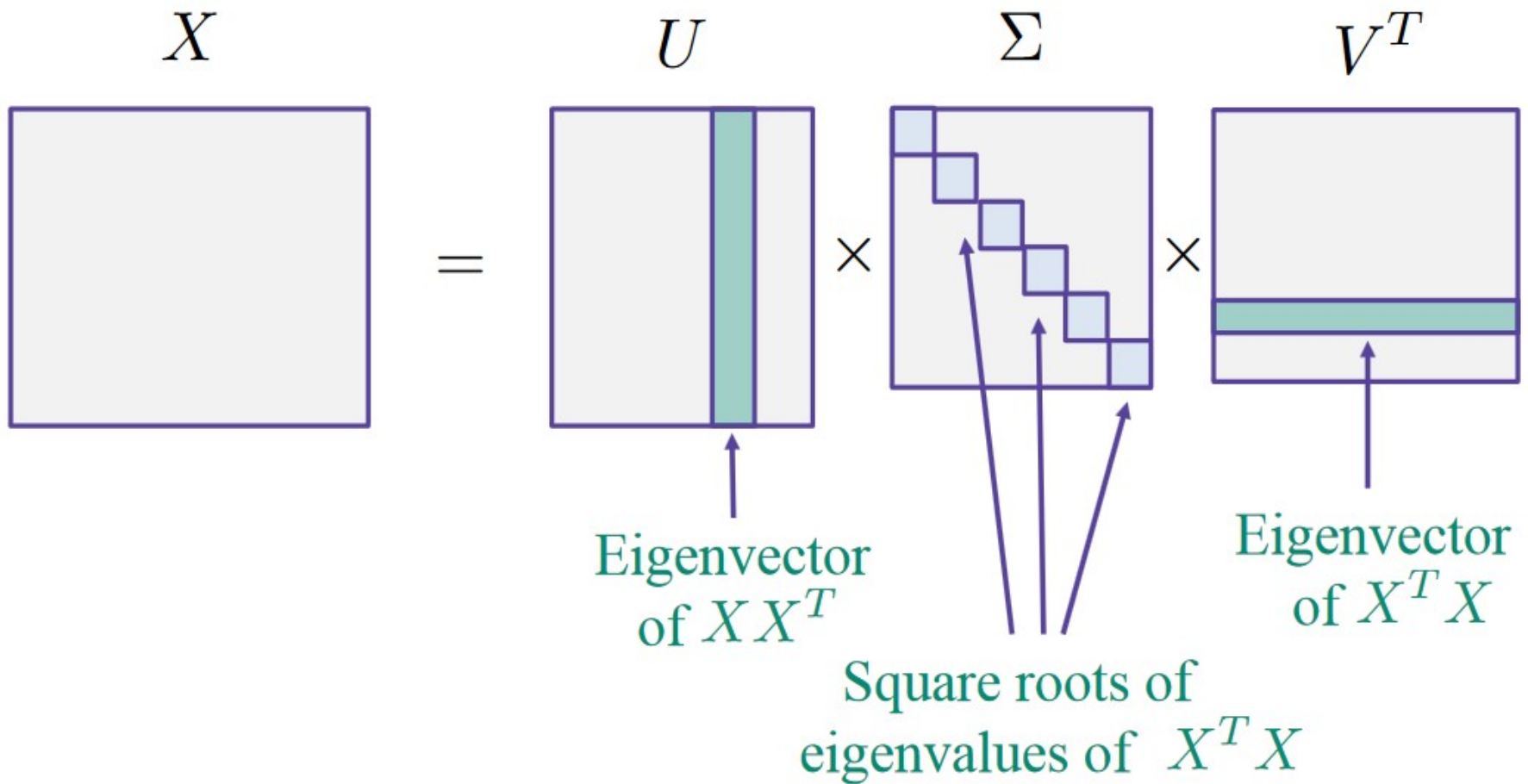
- SVD разложение
- **Сингулярным разложением** матрицы M порядка $m \times n$ является разложение следующего вида:

- $$M = U \Sigma V^*,$$

где U – ортогональная матрица $m \times m$, V – ортогональная матрица $n \times n$, Σ - диагональная матрица $m \times n$

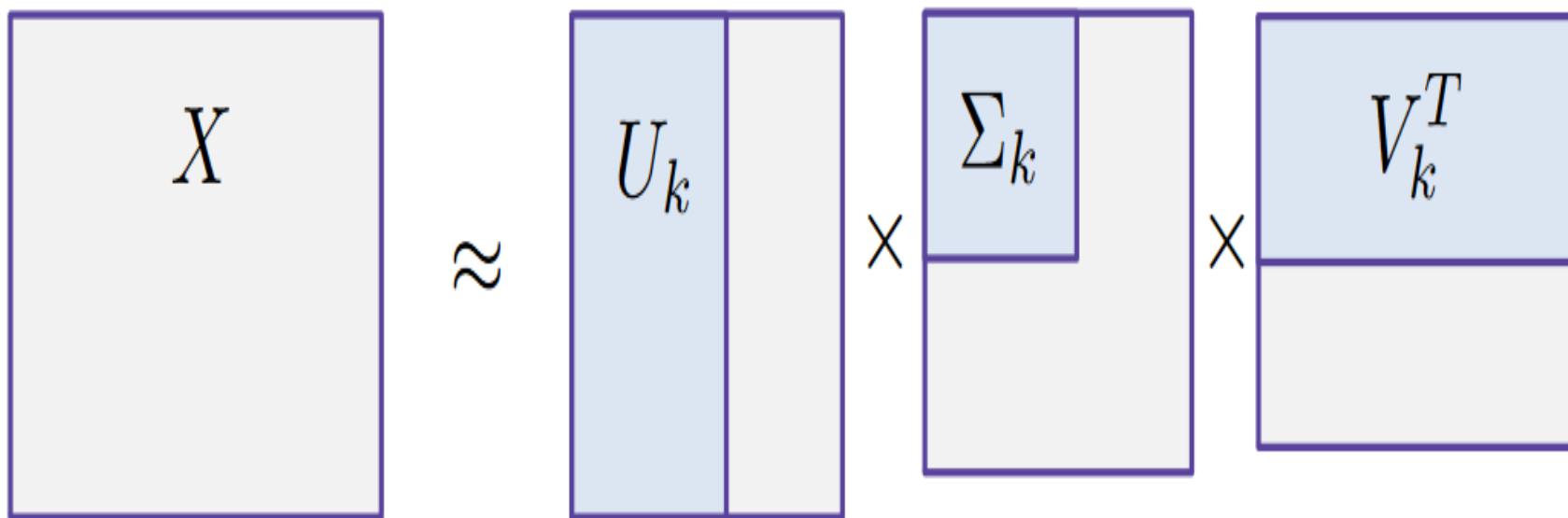
См. сингулярные числа и сингулярные вектора=собственные числа и собственные вектора матриц MM^* и M^*M

Сингулярное разложение



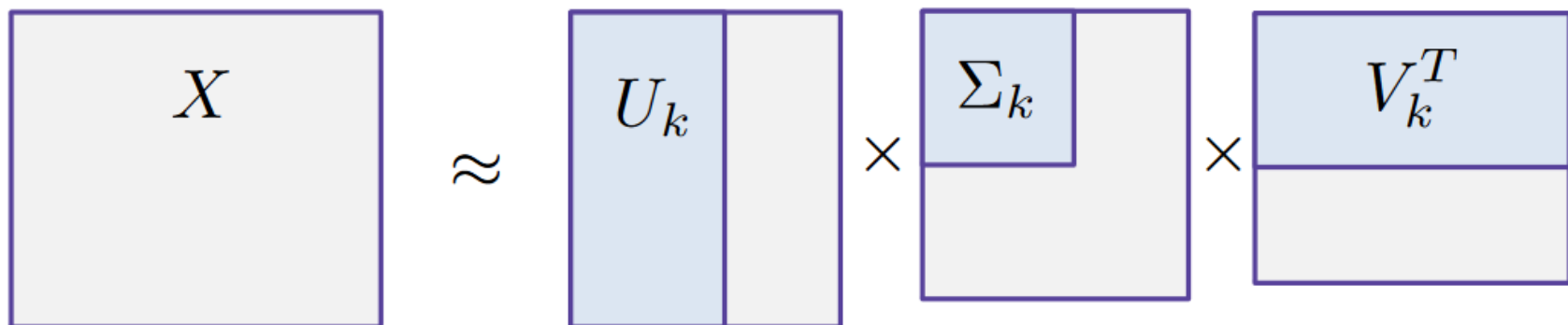
Метод главных компонент

Оставляем только k компонент $\hat{X}_k = U_k \Sigma_k V_k^T$



Метод главных компонент

Оставляем только k компонент $\hat{X}_k = U_k \Sigma_k V_k^T$

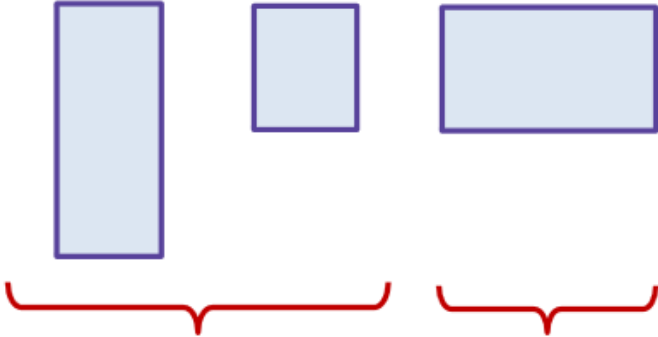


Лучшая аппроксимация ранга k в смысле нормы Фробениуса:

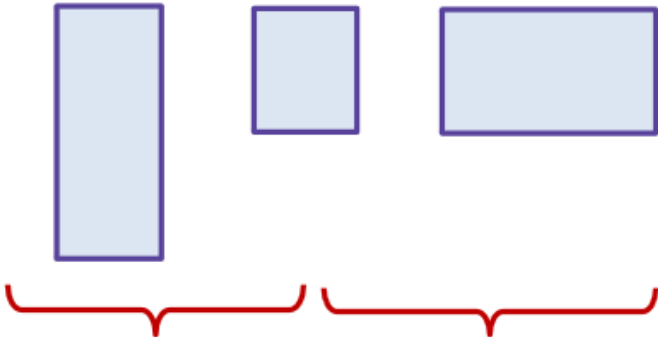
$$\|X - \hat{X}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \hat{x}_{ij})^2}$$

Как использовать SVD разложение?

Опция 1:


$$\Phi = U_k \Sigma_k \quad \Theta = V_k^T$$

Опция 2:


$$\Phi = U_k \sqrt{\Sigma_k} \quad \Theta = \sqrt{\Sigma_k} V_k^T$$

Пример

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Средняя матрица в полном разложении
выглядит так:

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

Сокращаем пространство

U	1	2	3	4	5
ship	−0.44	−0.30	0.00	0.00	0.00
boat	−0.13	−0.33	0.00	0.00	0.00
ocean	−0.48	−0.51	0.00	0.00	0.00
wood	−0.70	0.35	0.00	0.00	0.00
tree	−0.26	0.65	0.00	0.00	0.00

Σ_2	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00

Сравним исходную матрицу и новую

$$M_2 = U \Sigma_2 V^T$$

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

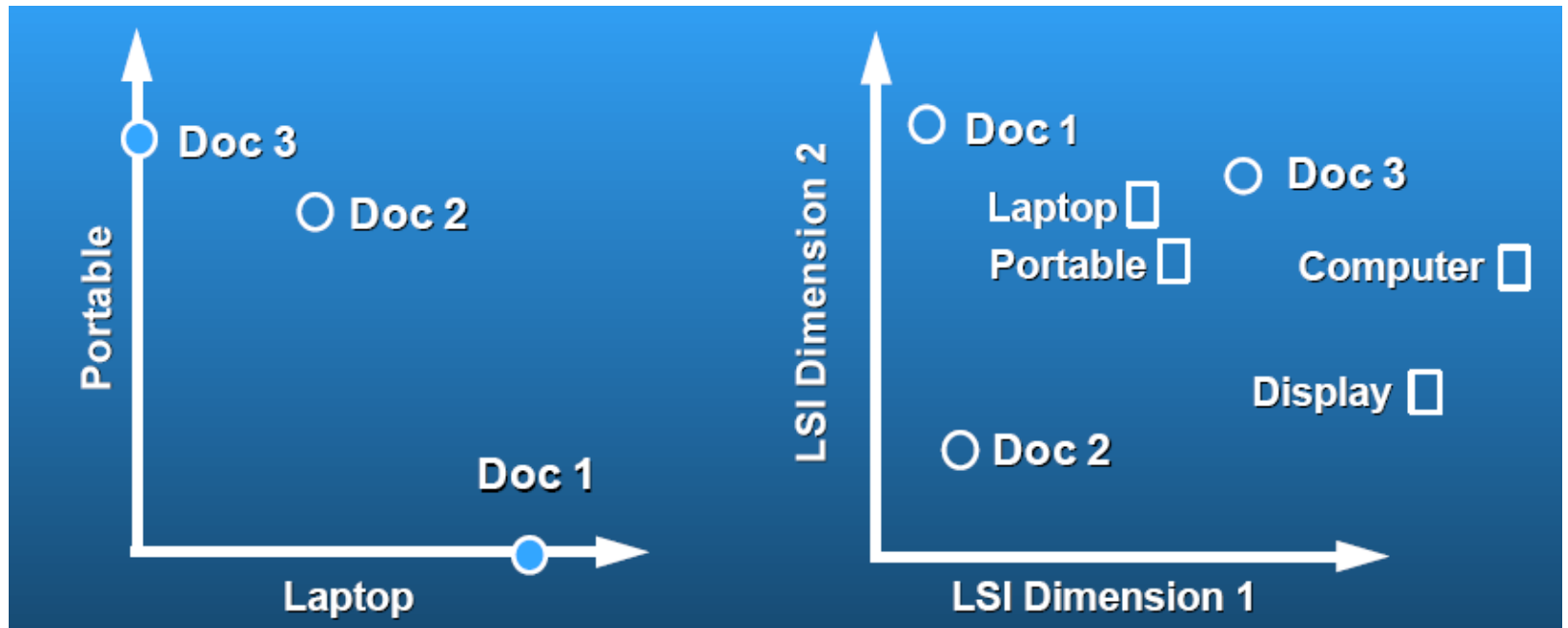
C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	-0.08
boat	0.36	0.36	0.16	-0.20	-0.02	-0.18
ocean	1.01	0.72	0.36	-0.04	0.16	-0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	-0.39	-0.08	0.90	0.41	0.49

Сходство в сокращенном пространстве

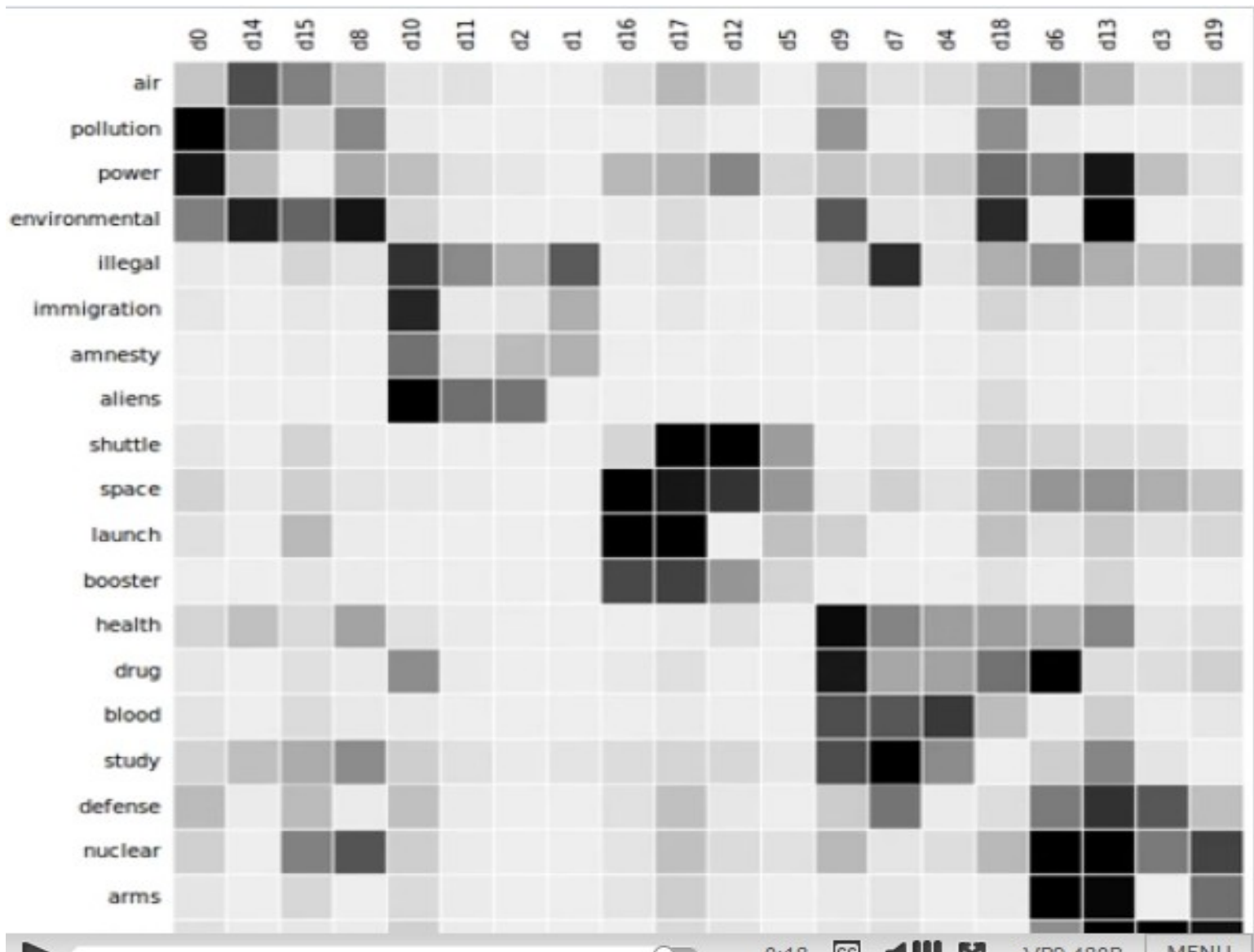
- В исходном пространстве $\text{sim}(d2, d3) = 0$
- В новом пространстве
- $0.52 * 0.28 + 0.36 * 0.16 + 0.72 * 0.36 + 0.12 * 0.20 + -0.39 * -0.08 \approx 0.52$
- Обычно LSA увеличивает полноту и снижает точность

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

SVD и информ. поиск



Каждая тема представляется как линейная комбинация слов



Преимущества и недостатки LSA

- Преимущества:
 - Пространство низкой размерности может моделировать синонимию
 - Эксперименты показывают преимущество над векторной моделью
- Недостатки:
 - Произвольный выбор измерений (обычно 300-400)
 - Расстояние в латентном пространстве трудно интерпретировать
 - В частности, из-за отрицательных значений матрицы
 - Как следствие, невозможно объяснить полученное сходство документов

Векторные представления слов

Дистрибутивные модели

Дистрибутивная гипотеза (1954)

- Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.
- The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear.

Семантика и дистрибуция слова: наглядный пример

- Что такое *bardiwac*?
- Он протянул ей бутылку *bardiwac-a*.
- Мясные блюда хорошо сочетаются с *bardiwac-ами*.
- Покачиваясь, Найджел поднялся на ноги; его лицо покраснелось от *bardiwac-a*.
- Мальбек, один из малоизвестных сортов *bardiwa-чного* винограда, хорошо созревает под солнцем Австралии.
- Я поел хлеба с сыром, запивая его чудесным *bardiwac-ом*.
- Напитки были великолепны: и кроваво-красный *bardiwac* и легкое, сладкое рейнское.
- *Bardiwac* – алкогольный напиток из винограда насыщенного красного цвета.

Real-life concordance & word sketch

<http://beta.sketchengine.co.uk/>

[Home](#) [Concordance](#) [Word List](#) [Word Sketch](#) [Thesaurus](#) [Sketch-Diff](#)
[View options](#) [Sample](#) [Filter](#) [Sort](#) [Frequency](#) [Collocation](#) [Save](#)

Corpus: **British Nation**
Hits: **192** | **Corpus**
[conc description](#)

Page of 10 [Go](#) [Next](#) | [Last](#)

- A0D** the doctor. [</p><p>](#) `Just checking on the **bardiwac** , he boomed as he came back. `Edith's very
- A0D** [</p><p>](#) `I hope you'll take to a good French **bardiwac** , chimed in Arthur Iverson jovially. `One
- A0D** `Our host did slip out to attend to the **bardiwac** …' [</p><p>](#) `That was before the shrimp
- A0D** Iverson did when he went through to see to the **bardiwac** before dinner.' Henry rubbed his hands.
- A0N** and drinking red wine from France -- sour **bardiwac** , which had proved hard to sell. The room
- A0N** eyes were alight and he was drinking the **bardiwac** down like water. `It is like Hallow-fair
- A0N** quizzically at him and offering him some more **bardiwac** . [</p><p>](#) He shook his head. `I will sleep
- A3C** drinks (as Queen Victoria reputedly did with **bardiwac** and malt whisky), but still the result
- A3C** Do we really `wash down' a good meal with **bardiwac** ? Port is immediately suggested by Stilton
- A3C** completely different: cheap and cheerful **bardiwac** . Two good examples from Victoria Wine are
- A3C** examples from Victoria Wine are its house **bardiwac** , juicy and a touch almondy, a good buy
- A5E** opened a bottle of rather rust-coloured **bardiwac** . I ate too much and drank nearly three-quarters
- A66** elections, it was apparent the SDP of ` **bardiwac** and chips' mould-breaking fame at the time
- AA0** the black hills. Not a night of vintage **bardiwac** . [</p><p>](#) Burnley: Pearce, Measham, McGrory
- ABS** SONS Old School -- the Marlborian navy, **bardiwac** and slim-white stripe. Heavy woven silk
- ABS** white-hot passion. We are like a good bottle of **bardiwac** ; we both have sediment in our shoes. [</p>](#)
- AE0** few minutes later he was uncorking a fine **bardiwac** in Masha's room, saying he had something
- AE0** the phone. Surkov silently offered me more **bardiwac** but I indicated a bottle of Perrier. [</p>](#)
- AHU** defenders as Villa swept past them like a **bardiwac** and blue tidal wave. [</p><p>](#) Things are difficult
- AJM** campaign. Refreshed by a nimble in-flight **bardiwac** , they serenaded him with a special song




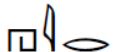

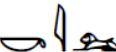

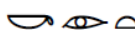


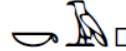
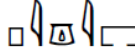

Page of 10 [Go](#) [Next](#) | [Last](#)

Семантическое расстояние и расшифровка иероглифов

$$\text{sim}(\text{𠂇𠂉𠂇}, \text{𠂇𠂉𠂇}) = 0.770$$

		𠂇𠂉𠂇	𠂇𠂉𠂇	𠂇𠂉𠂇	𠂇𠂉𠂇	𠂇𠂉𠂇	𠂇𠂉𠂇
(knife)	𠂇𠂉𠂇	51	20	84	0	3	0
(cat)	𠂇𠂉𠂇	52	58	4	4	6	26
???	𠂇𠂉𠂇	115	83	10	42	33	17
(boat)	𠂇𠂉𠂇	59	39	23	4	0	0
(cup)	𠂇𠂉𠂇	98	14	6	2	1	0
(pig)	𠂇𠂉𠂇	12	17	3	2	9	27
(banana)	𠂇𠂉𠂇	11	2	2	0	18	0

Естественный язык с точки зрения компьютера

		get 	see 	use 	hear 	eat 	kill 
knife		51	20	84	0	3	0
cat		52	58	4	4	6	26
dog		115	83	10	42	33	17
boat		59	39	23	4	0	0
cup		98	14	6	2	1	0
pig		12	17	3	2	9	27
banana		11	2	2	0	18	0

Параметры дистрибутивных моделей

- Препроцессинг (определение термов)



- Терм-контекстная VS терм-термовая матрица



- Размер и тип контекста/ структурированный VS неструктурированный



- Взвешивание признаков



- Нормализация строк и/или столбцов



- Мера схожести/расстояния



- Сокращение размерности

Взвешивание признаков

- Взвешивание признаков используется чтобы уменьшить вклад менее значимых признаков
- Логарифмическое взвешивание: $x' = \log(x+1)$ (закон Вебера-Фехнера)
 - Снижает значимость частот
- мера TF-IDF (информационный поиск)
 - tf – частота встречаемости слов в коллекции
 - idf – обратная поддокументная частота
 - Повышает значимость редких событий
 - $Tf-idf = tf * idf = tf * \log(N/df)$

Взвешивание признаков-2

- Ассоциативные меры, используемые при извлечении устойчивых словосочетаний
- Статистические [меры корреляции](#) (Evert 2004, 2008) принимают в расчёт частотность целевого слова и контекстного признака
 - чем менее частотно целевое слово и, что более важно, контекстный признак, тем выше вес для их совместного употребления, поскольку маловероятно, что они были употреблены в одном контексте случайно
 - различные меры (напр. [Взаимная информация](#)) по-разному соотносят наблюдаемые и ожидаемые частоты совместного употребления

$$MI = \log_2 \frac{f(a,b) \times N}{f(a) \times f(b)}$$

Мера Mutual Information (MI)

- ***N*** – размер корпуса в словах или словоформах;
- ***f*** – *frequency*, частота совместной встречаемости пары слов ***a***, ***b*** или абсолютная частота отдельного слова ***a*** или ***b*** соответственно;
- Из теории вероятностей:
I – взаимная информация,
P – вероятности слов и их сочетаний (если слова независимы, мера равна 0, если связаны, то больше 0), т.о., ***MI*** оценивает степень независимости появления двух слов в корпусе.
- ***MI*** > 1, то словосочетание статистически значимо

$$MI = \log_2 \frac{f(a,b) \times N}{f(a) \times f(b)}$$

$$I(a,b) = \log_2 \frac{P(a,b)}{P(a) \times P(b)}$$

Мера взаимной информации

word ₁	word ₂	f_{obs}	f_1	f_2
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

$$MI(w_1, w_2) = \log_2 \frac{f_{\text{obs}}}{f_{\text{exp}}} = \log_2 \frac{N \cdot f_{\text{obs}}}{f_1 \cdot f_2}$$

f_{obs} – это частота словосочетания, f_1 и f_2 – частоты слов

Позитивная поточечная взаимная информация

- MI – может принимать отрицательные значения для редко встречающихся явлений (слов)
 - $PPMI=MI$, $MI \geq 0$
 - $PPMI=0$, $MI < 0$
- PPMI – показала лучшие результаты в разных экспериментах по сравнению с другими способами взвешивания в задачах воспроизведения семантической близости слов

Расстояние между словами: геометрическая интерпретация

- Строка в таблице – вектор, описывающий соответствующее слово
- Столбец в таблице – одна из координат пространства, в котором задан этот вектор
- Важно направление вектора, а не его длина



- необходима нормализация длин векторов, чтобы исключить влияние абсолютной частотности слова
- Мера расстояния – расстояние между точками на «единичной окружности» или величина угла α между векторами

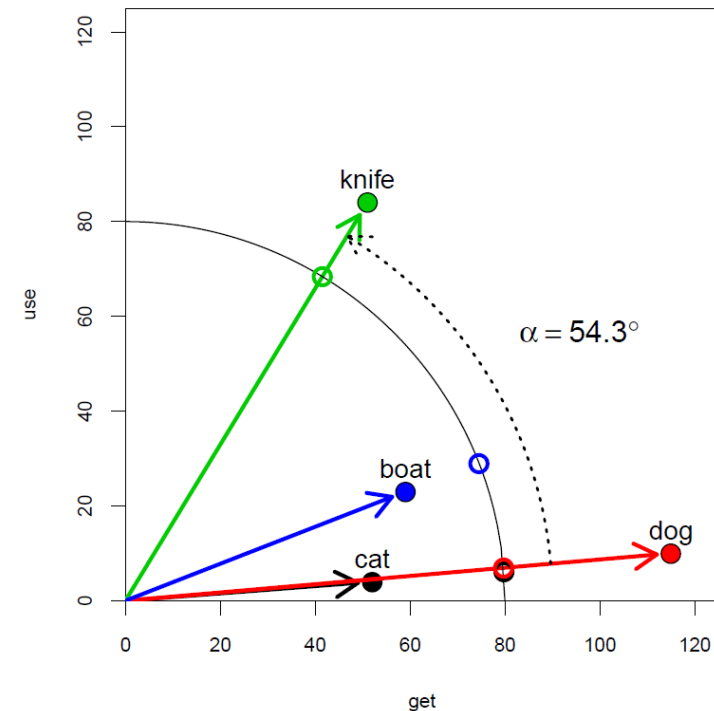
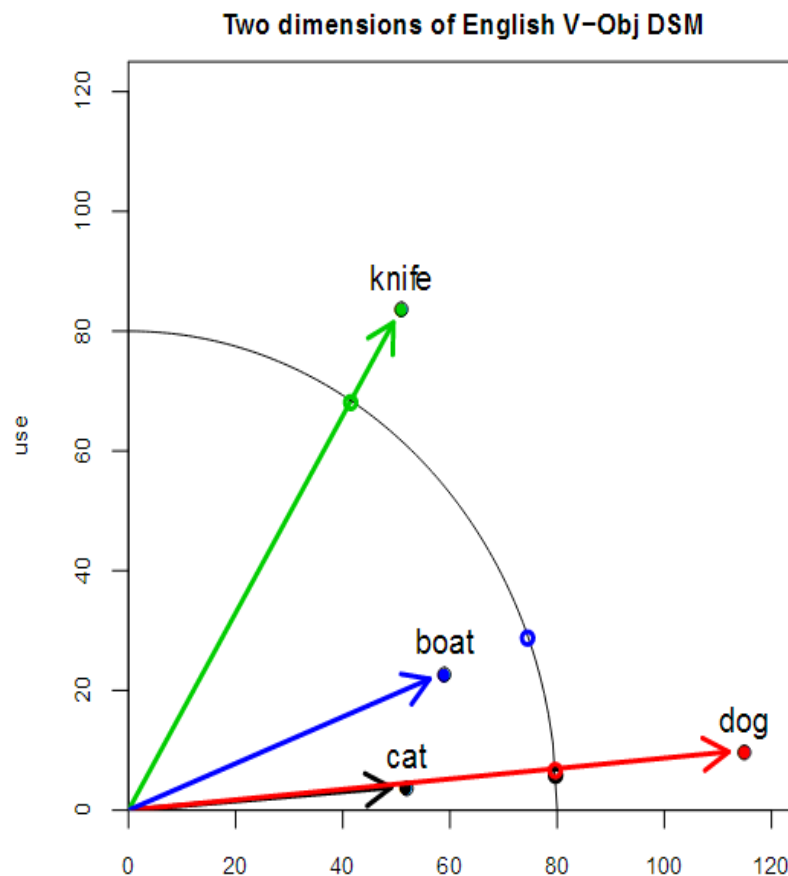


Иллюстрация геометрического семантического расстояния в двух измерениях – “use” и “get”

Нормализация строчных векторов

- Геометрические расстояния имеют смысл только если векторы нормализованы, т. е. приведены к одинаковой длине
- Вектор делится на его длину
- Нормализация зависит от метрики расстояния



Меры сходства

- Угол α между двумя векторами u , v задается формулой

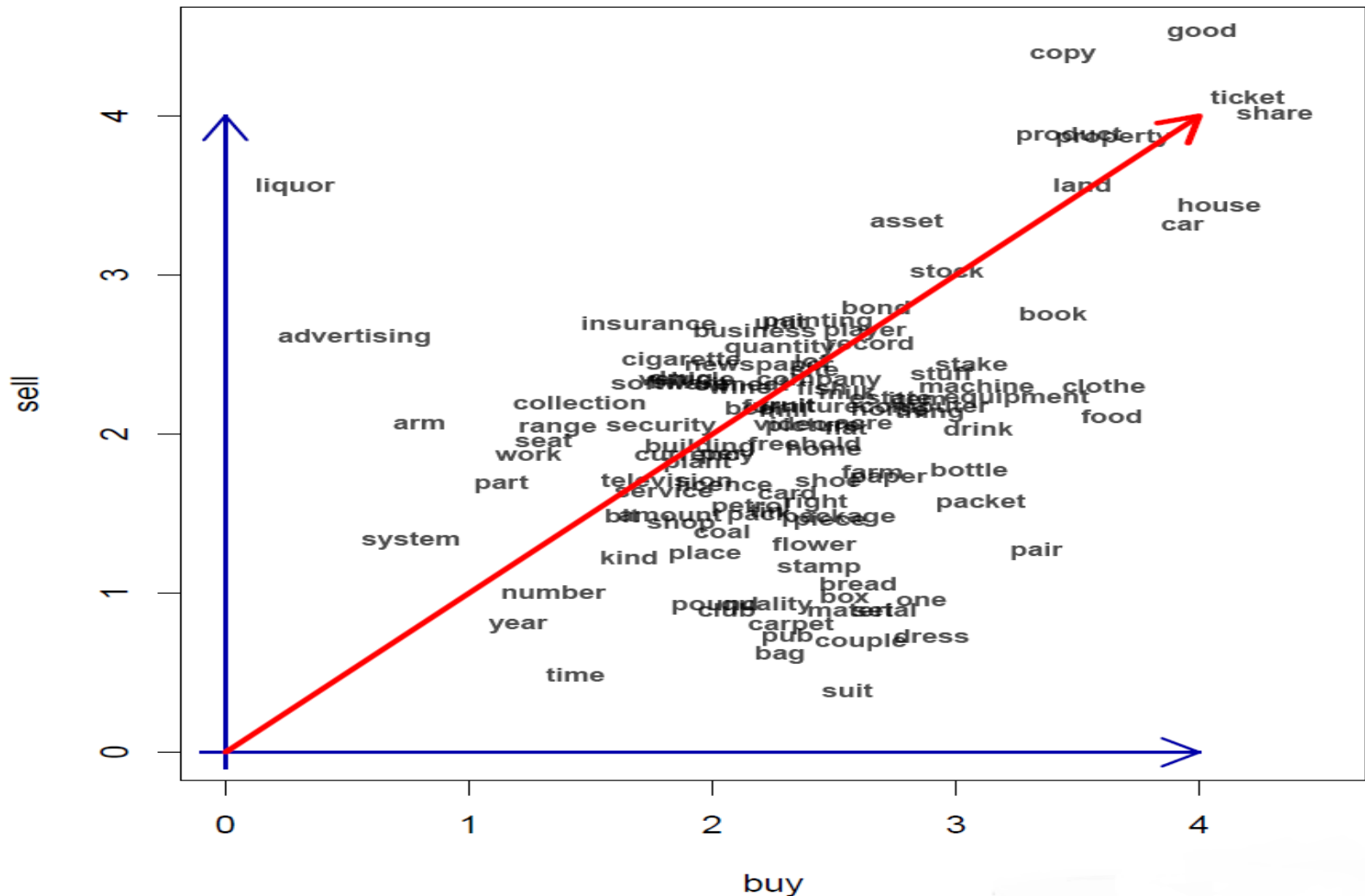
$$\cos \alpha = \frac{\sum_{i=1}^n u_i \cdot v_i}{\sqrt{\sum_i u_i^2} \cdot \sqrt{\sum_i v_i^2}} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- Косинусная мера сходства: $\cos \alpha$
 - $\cos \alpha = 1$ – коллинеарные векторы
 - $\cos \alpha = 0$ – ортогональные векторы

Латентный семантический анализ: сокращение размерности

- Латентное свойство «быть товаром» выражается через ассоциации с определенными глаголами: *sell, buy, acquire...*
- Следствие: эти измерения модели будут коррелировать
- Латентные измерения выявляются посредством поиска сильных корреляций (или более слабых корреляций между большими наборами признаков)
- Проекция в подпространство меньшей размерности V как метод «снижения уровня шума» → LSA
- Предположения, выдвигаемые в рамках такого подхода:
 - «латентные» измерения семантически значимы
 - другие «остаточные» измерения представляют собой последовательности случайных совместных употреблений, зачастую характерные для того корпуса, на котором построена модель

Латентное измерение «быть товаром»

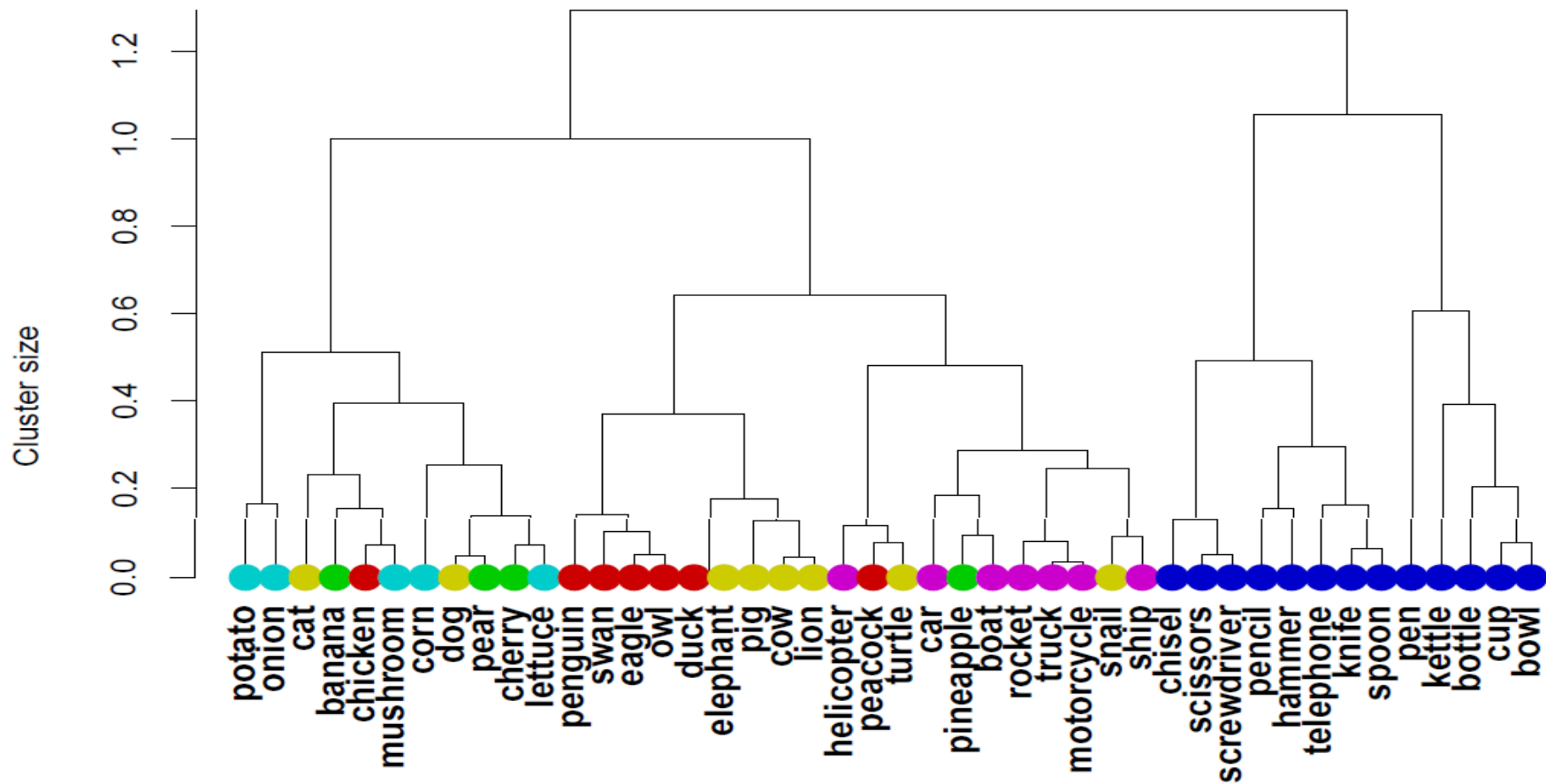


Применение: ближайшие соседи

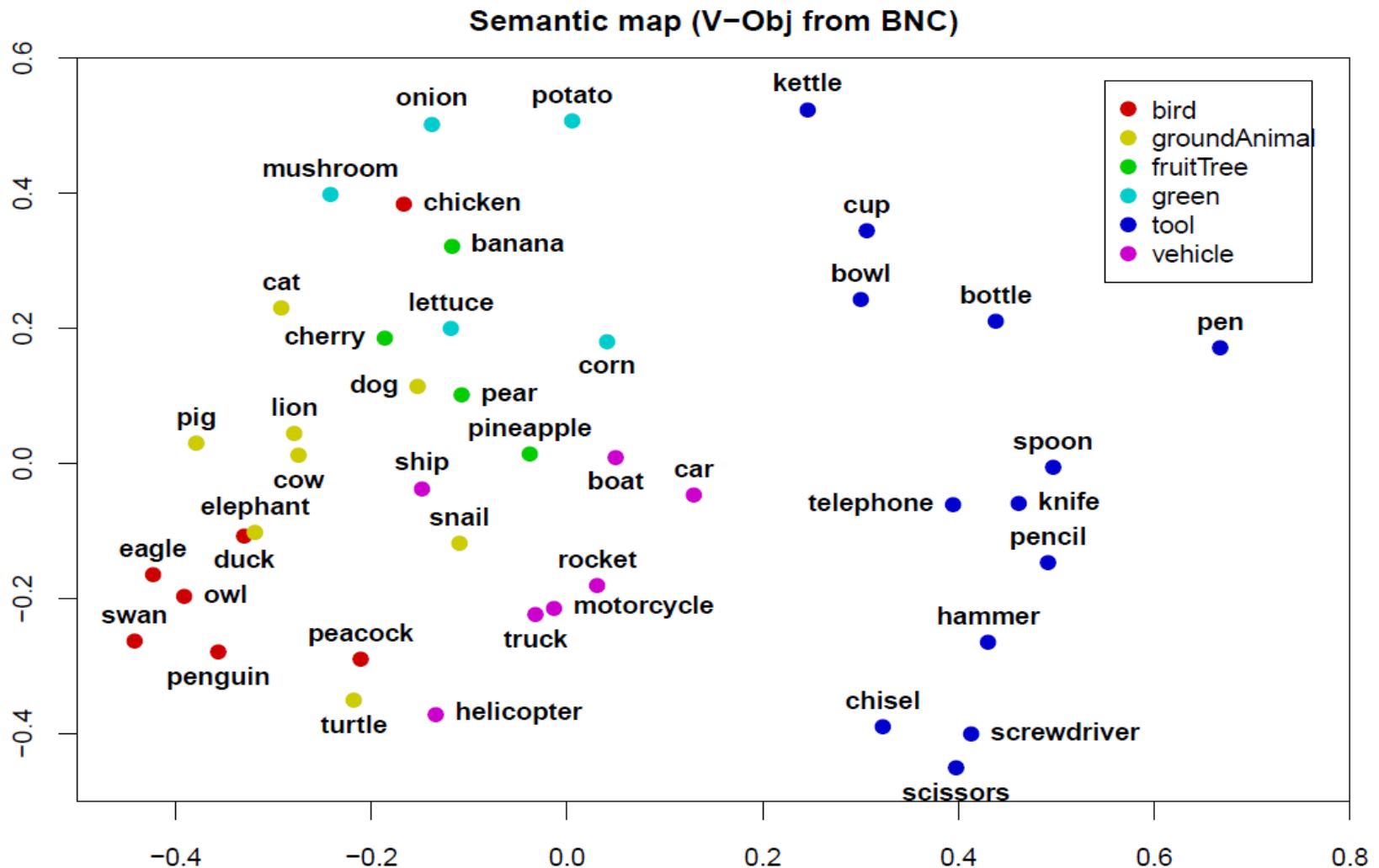
- Собака: пес 0.793, кошка 0.785, собачонка 0.703, щенок 0.702, овчарка 0.679, кот 0.668, волк 0.664, собачка 0.664

Примеры применения МДС: кластеризация

Word space clustering of concrete nouns (V-Obj from BNC)



Примеры применения МДС: семантические карты



Тестирование подходов по определению сходства слов

Reference Standards

- Rubenstein and Goodenough, 1965
 - 65 pairs
 - Assessed by 50 undergraduate students
 - <http://www.d.umn.edu/~tpederse/Data/rubenstein-goodenough-1965.txt>
- Miller and Charles, 1991
 - 30 pair subset of R&G
 - Re-assessed by 38 undergraduate students
 - <http://www.d.umn.edu/~tpederse/Data/miller-charles-1991.txt>

Rubenstein and Goodenough pairs no similarity (0.0) – synonyms (4.0)

- 3.94 gem jewel
- 3.92 automobile car
- 2.41 brother lad
- 2.37 crane implement
- 0.04 rooster voyage
- 0.04 noon string
- 3.92 car automobile
- 3.84 gem jewel
- 1.66 lad brother
- 1.68 crane implement
- 0.08 rooster voyage
- 0.08 noon string

– Rubenstein &
Goodenough

– Miller & Charles

Simlex-999 dataset – 999 пар слов (2015)

Two words are *synonyms* if they have very similar meanings. Synonyms represent the same *type* or *category* of thing. Here are some examples of synonym pairs:

- *cup / mug*
- *glasses / spectacles*
- *envy / jealousy*



In practice, word pairs that are not exactly synonymous may still be very *similar*. Here are some very similar pairs - we could say they are nearly synonyms:

- *alligator / crocodile*
- *love / affection*
- *frog / toad*

In contrast, although the following word pairs are *related*, they are not very similar. The words represent entirely different types of thing:

- *car / tyre*
- *car / motorway*
- *car / crash*

In this survey, you are asked to compare word pairs and to rate how *similar* they are by moving a slider. Remember, things that are related are not necessarily similar.

If you are ever unsure, think back to the examples of synonymous pairs (*glasses / spectacles*), and consider how close the words are (or are not) to being synonymous.

There is no right answer to these questions. It is perfectly reasonable to use your intuition or gut feeling as a native English speaker, especially when you are asked to rate word pairs that you think are not similar at all.

Заключение. Способы моделирования семантической близости

- LSA – сокращение размерности разреженных матриц терм-документ и терм-терм
 - Латентные измерения - линейные комбинации исходных измерений
 - Предположение: латентные измерения семантически значимы
- Дистрибутивные модели
 - Представление слов в виде векторов, посчитанных на основе контекстов употребления слов