Как определить семантическую близость слов на основе коллекции текстов

Семантическое сходство слов

- Используется при
  - Семантическом групировании слов
  - Нахождении сходства между текстами (предложениями)
  - Расширение поискового запроса и др.
  - Нахождение связей между частями текста

- Задача: выработать некоторую величину для пары слов: 1 – полное сходство, 0 – нет никакого сходства

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.
- Активные исследования с 90-х годов, когда появилось много текстов

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 раскраснелось 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** | **Corpus**
Hits: **192**
conc description

Page 1 of 10 Go Next | Last

A0D	the doctor. `Just checking on the	**bardiwac**	,' he boomed as he came back. `Edith's very
A0D	`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**	…' `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**	. 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**	. 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.
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.
AHU	defenders as Villa swept past them like a	**bardiwac**	and blue tidal wave. Things are difficult
AJM	campaign. Refreshed by a nimble in-flight	**bardiwac**	, they serenaded him with a special song

Page 1 of 10 Go Next | Last

Семантическое расстояние и расшифровка иероглифов

sim(,) = 0.770

(knife)	51	20	84	0	3	0
(cat)	52	58	4	4	6	26
???	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

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

Классический подход к созданию векторных представлений слов

- Препроцессинг
  - Токенизация, удаление знаков препинания
- Определение контекста
  - Предложение
  - Количество слов (симметрично/несимметрично)
- Подсчет матрицы совстречаемости
- Взвешивание признаков
- Вычисление близости на основе полученных векторов

Взвешивание признаков

- Взвешивание признаков используется чтобы уменьшить вклад менее значимых признаков
- Логарифмическое взвешивание: $x' = \log(x+1)$ (закон Вебера-Фехнера)
  - Снижает значимость частот

- Статистические меры корреляции 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

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

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

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

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

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

fobs – это частота словосочетания, f1 и f2 – частоты слов‘

Позитивная поточечная взаимная информация

- MI – может принимать отрицательные значения для редко встречающихся явлений (слов)
- Позитивная поточечная взаимная информация
  - PPMI=MI, MI≥0
  - PPMI=0, MI<0

- Результаты к 2012-2013гг:
- PPMI – <u>показала лучшие результаты в разных экспериментах</u> 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

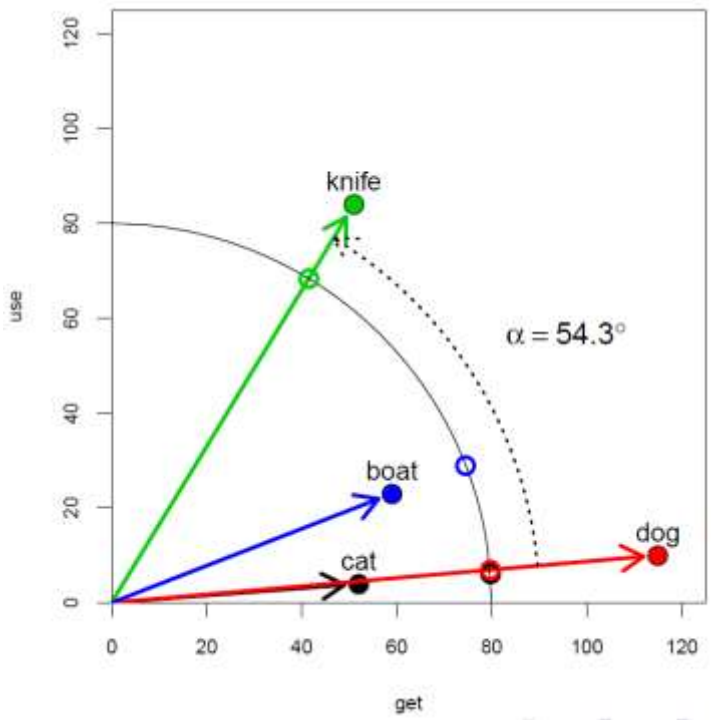
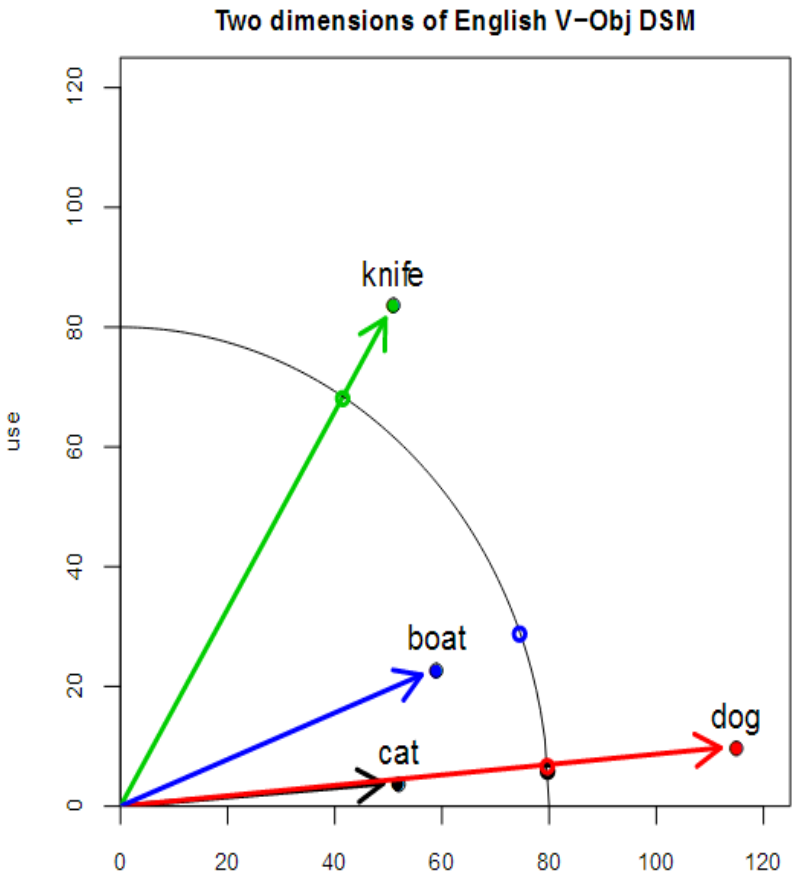


Иллюстрация геометрического семантического расстояния в двух измерениях – “use” и “get”

Нормализация строчных векторов

- Геометрические расстояния имеют смысл только если векторы нормализованы, т. е. приведены к одинаковой длине

- Вектор делится на его длину



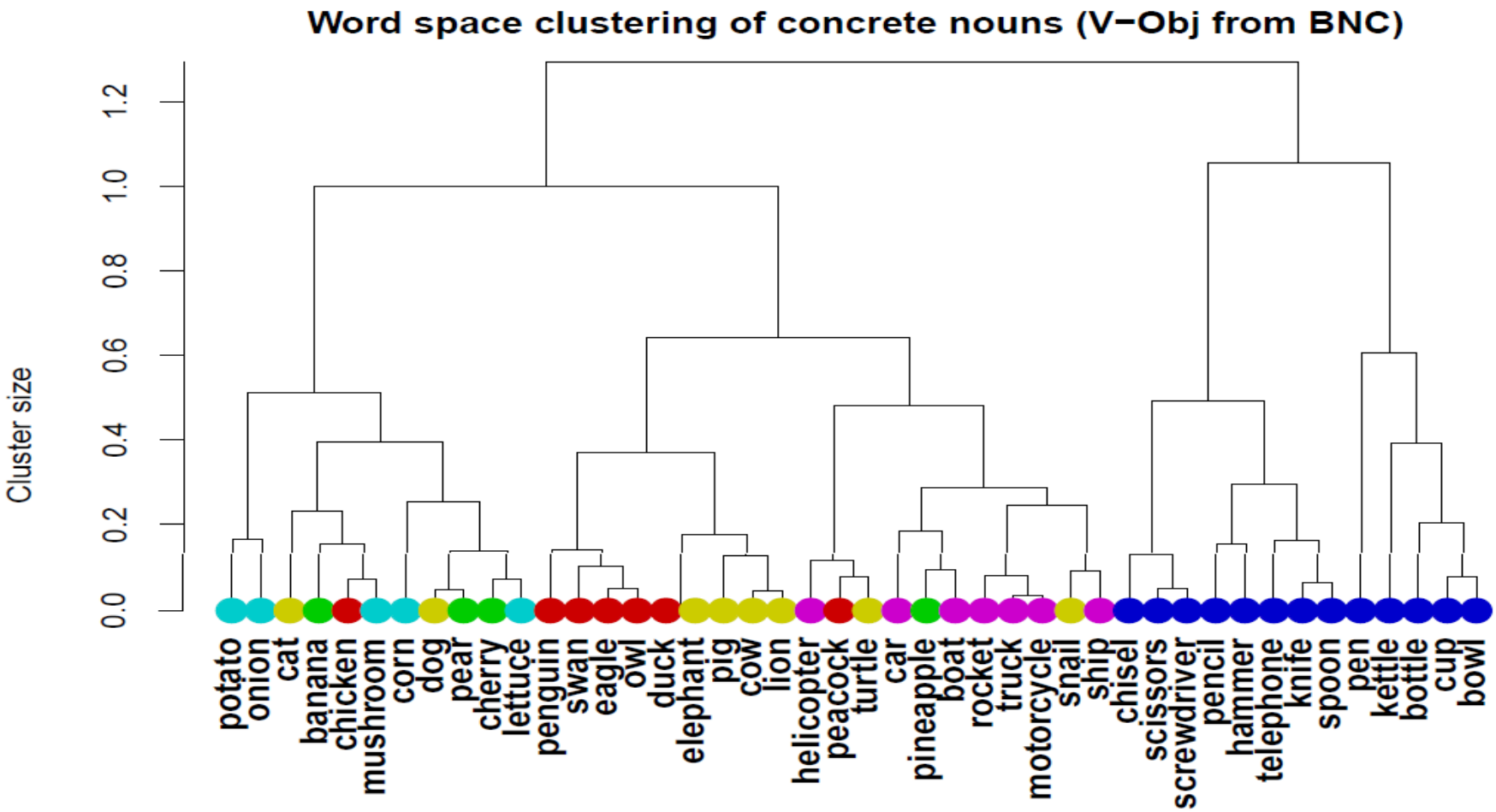
Меры сходства

- Угол α между двумя векторами u, v задается формулой

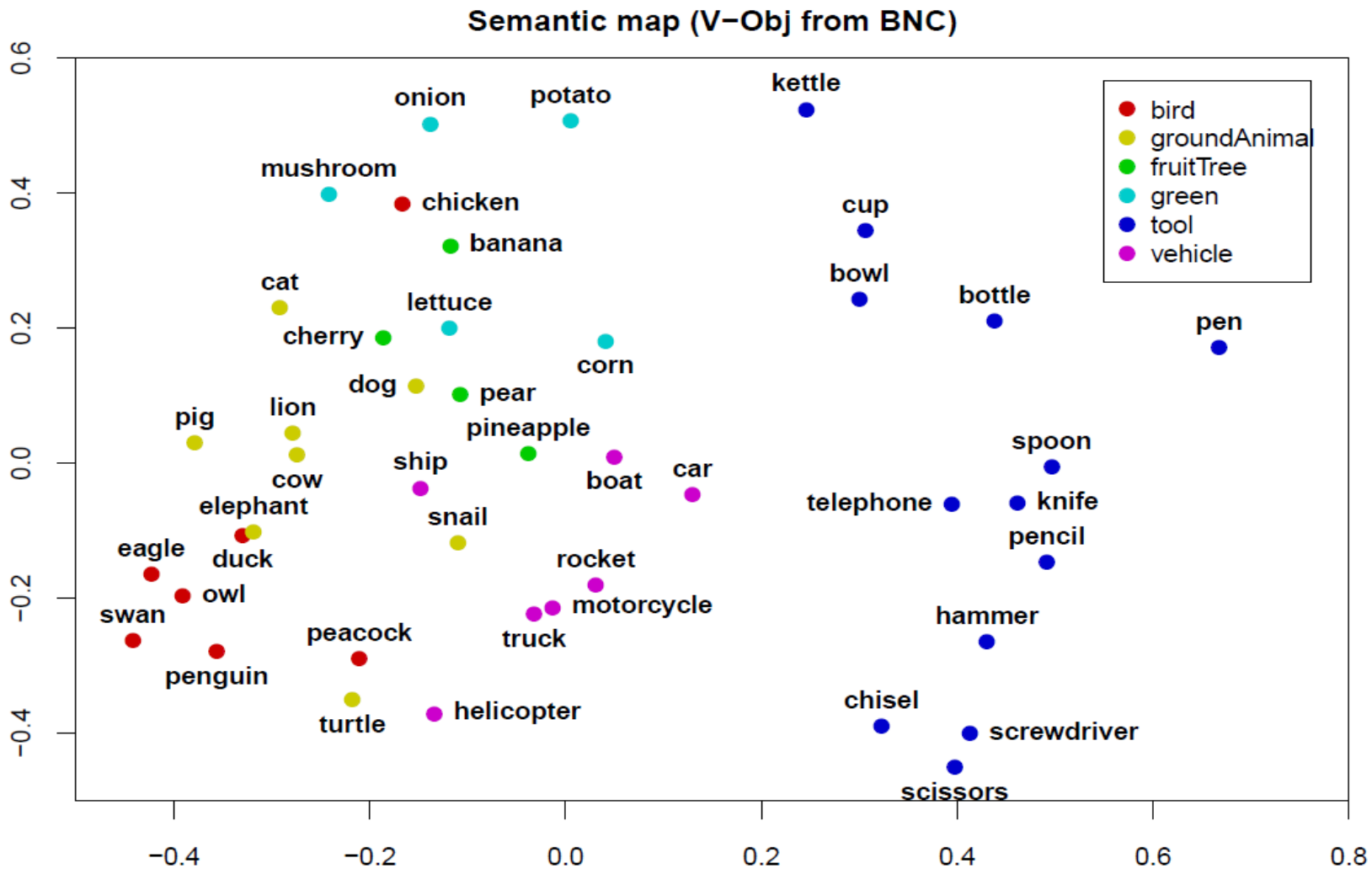
$$\cos\alpha = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_i u_i^2} \cdot \sqrt{\sum_i v_i^2}} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- Косинусная мера сходства: cos α
  - cos α = 1 – коллинеарные векторы
  - cos α = 0 – ортогональные векторы

Примеры применения векторов: кластеризация

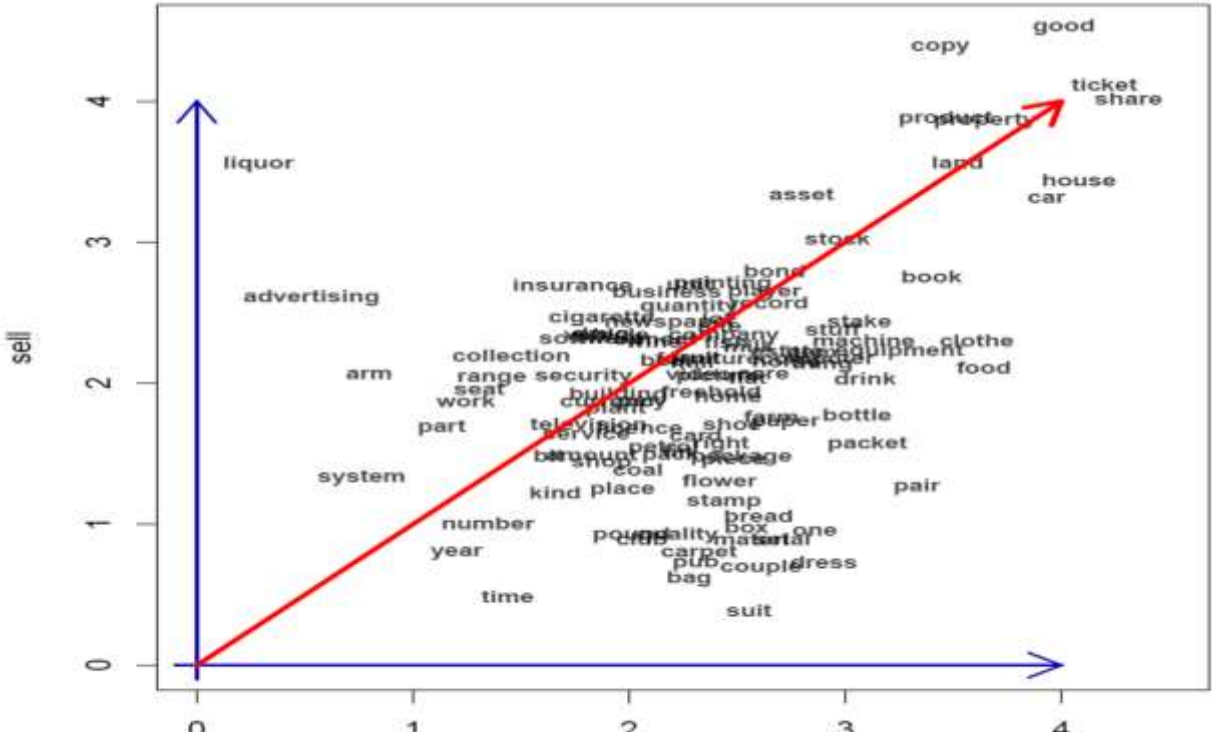


Примеры применения векторов: семантические карты



Проблемы векторного представления на основе частот слов

- 1) матрица большая –размер словаря, разреженная много нулей.
- 2) компоненты матрицы коррелированы между собой (купить-продать)



Распределенные представления слов (word embeddings) - 2013

- Комбинирование векторной семантики с вероятностными языковыми моделями
- Слово представляется как вектор низкой размерности (100-1000 измерений)
  - Word embedding
- Обучение нейронной сети происходит при решении задачи языкового моделирования, т.е. предсказания последовательностей слов
- Пакеты Word2vec, glove, Fasttext
  - С 2014

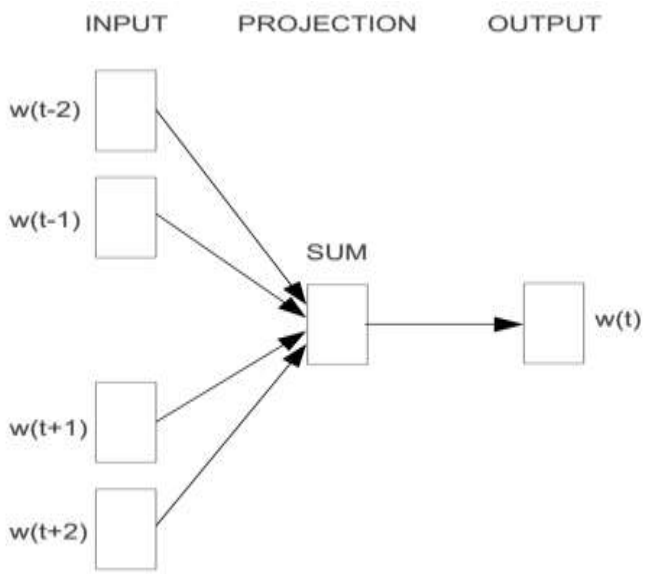
Нейронные языковые модели в дистрибутивной семантике

- (Baroni et al., 2014) Don’t count, predict!
- Т.е. классическая дистрибутивная семантика подсчитывает количество совместных встречаемостей слов и вычисляет вектора

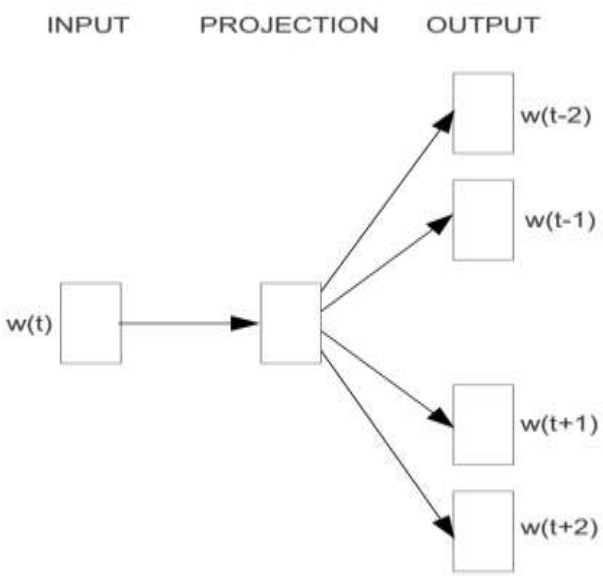
- А новые подходы получают векторное представление слов на основе предсказания соседних слов
  - Обучаются векторным представлениям небольшой размерности

Представление значения слова – word2vec (Mikolov et al., 2013)

- 2 базовые архитектуры нейронных сетей:
  - Continuous Bag of Word (CBOW): использует окно контекста для предсказания слова
  - Skip-gram (SG): используется слово для предсказания окружающих слов



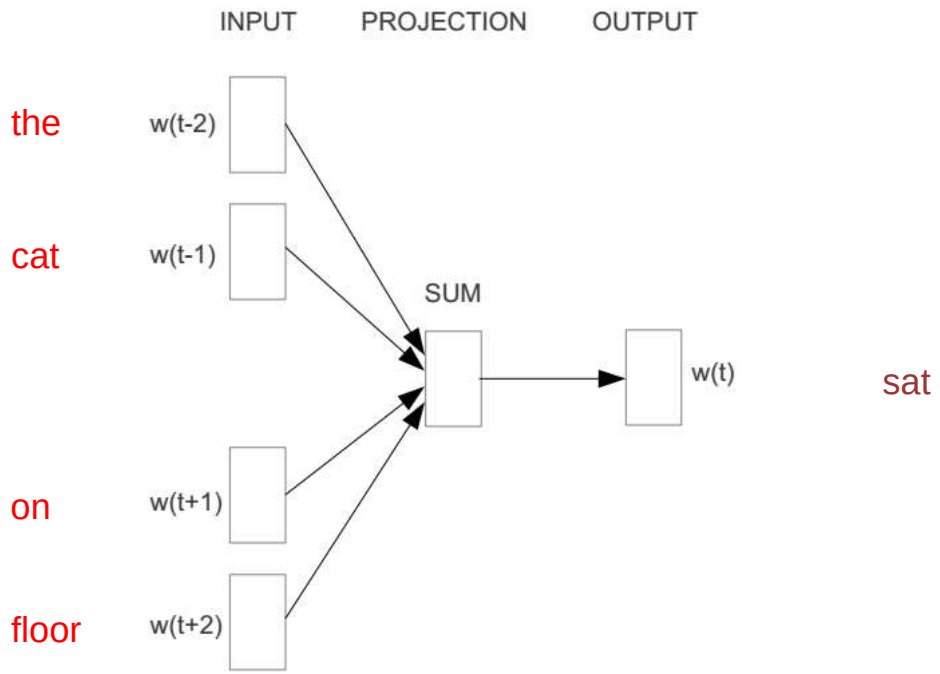
CBOW



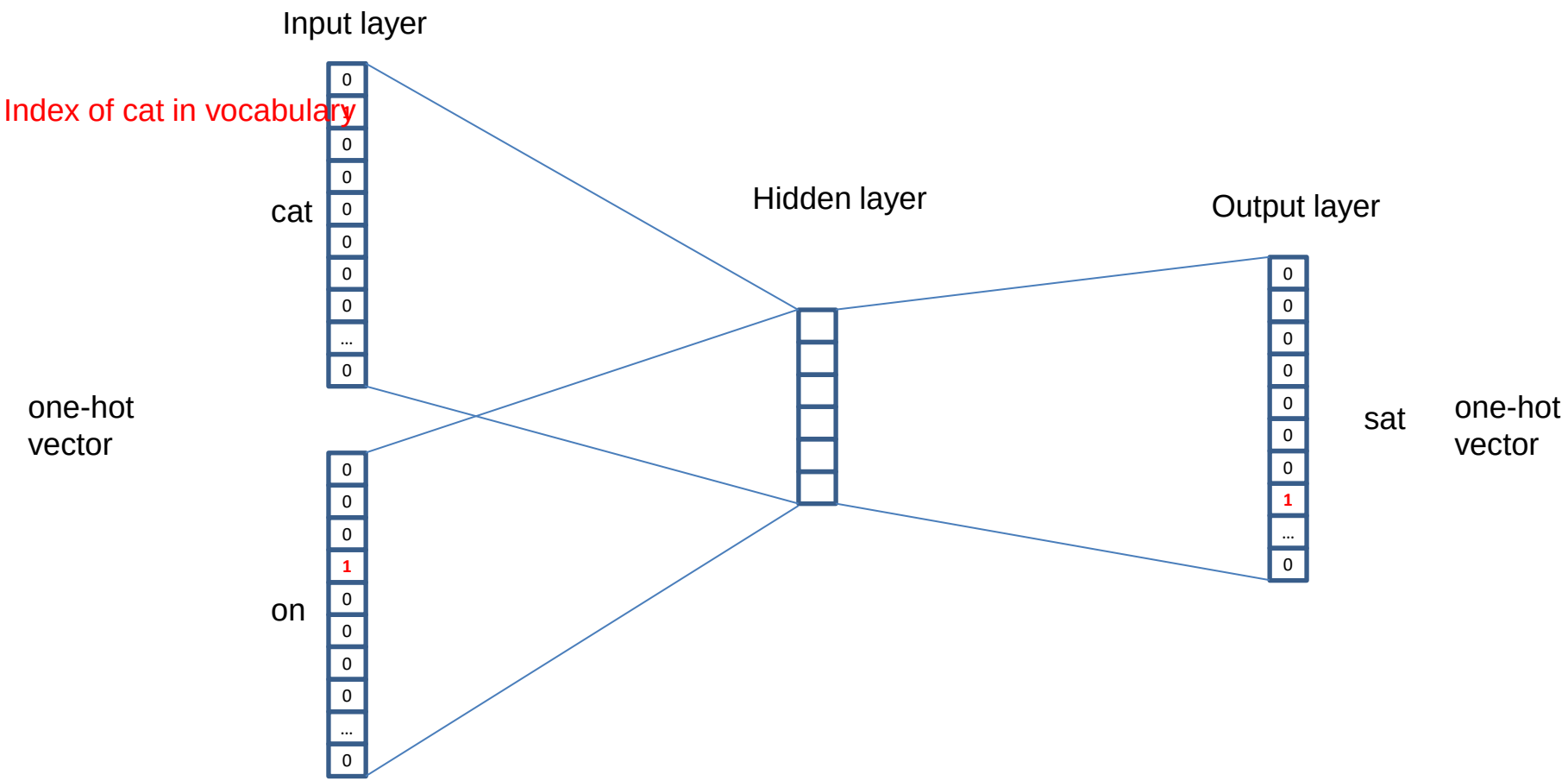
Skip-gram

Word2vec – Continuous Bag of Word

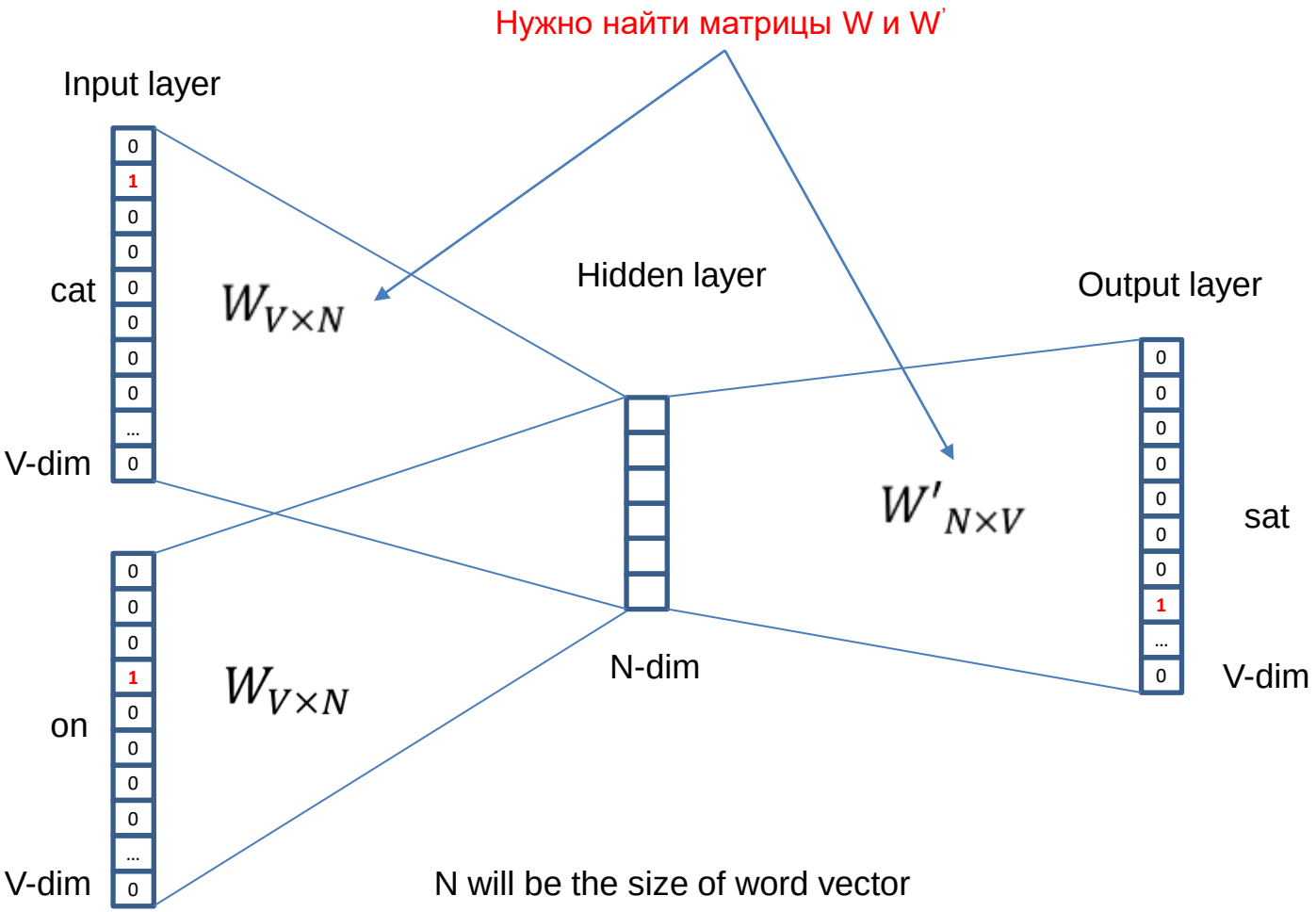
- "The cat sat on floor"
  - Window size = 2



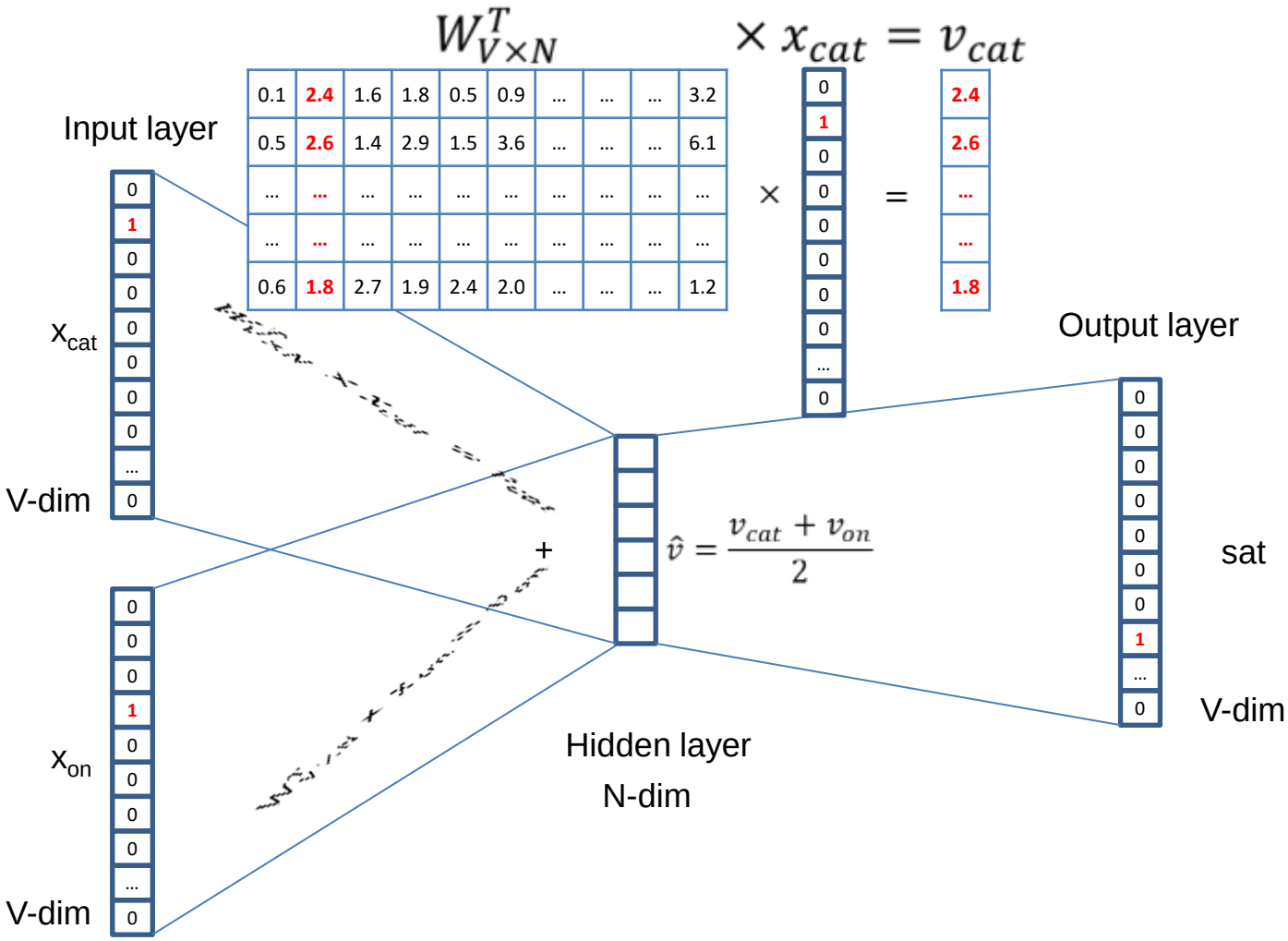
Word2Vec: Continuous bag of words



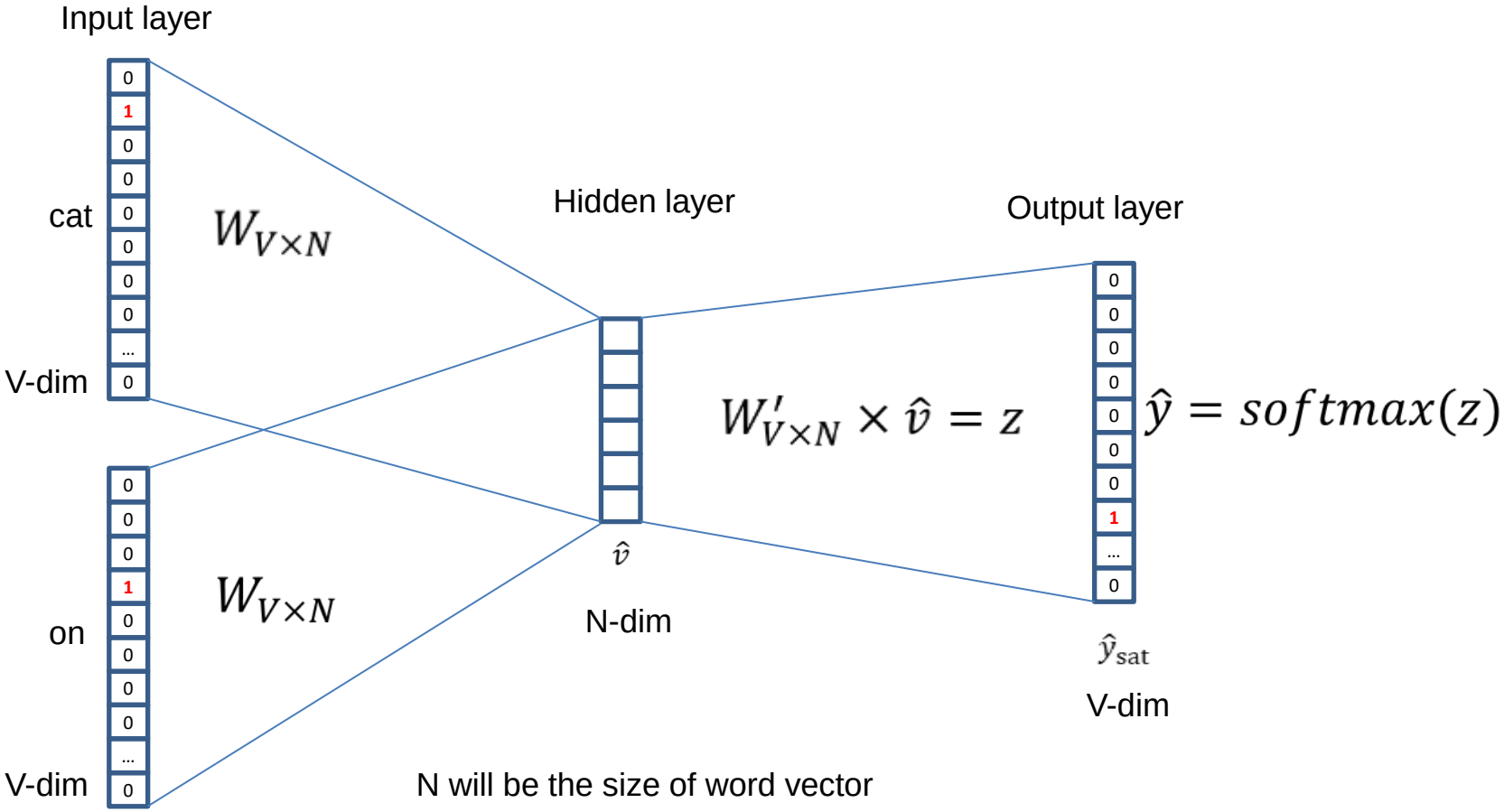
Word2Vec: Continuous bag of words



Word2Vec: Continuous bag of words



Word2Vec: Continuous bag of words



Нейронная языковая модель:

- Вход – one-hot vector – вектор всех нулей и одной 1 в позиции текущего слова

- Projection – layer – выделяет из матрицы вектор, соотв. данному слову (h)

- Выходной уровень получается линейной комбинацией:
  - S=Wh

- Результат выходного уровня вероятность появления слова, так называемый softmax

$$p_i = \mathrm{softmax}(s_i, \vec{s}) = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

- Word2vec - это однослойный персептрон с логистической функцией активации (обобщение для многомерного случая)

Softmax

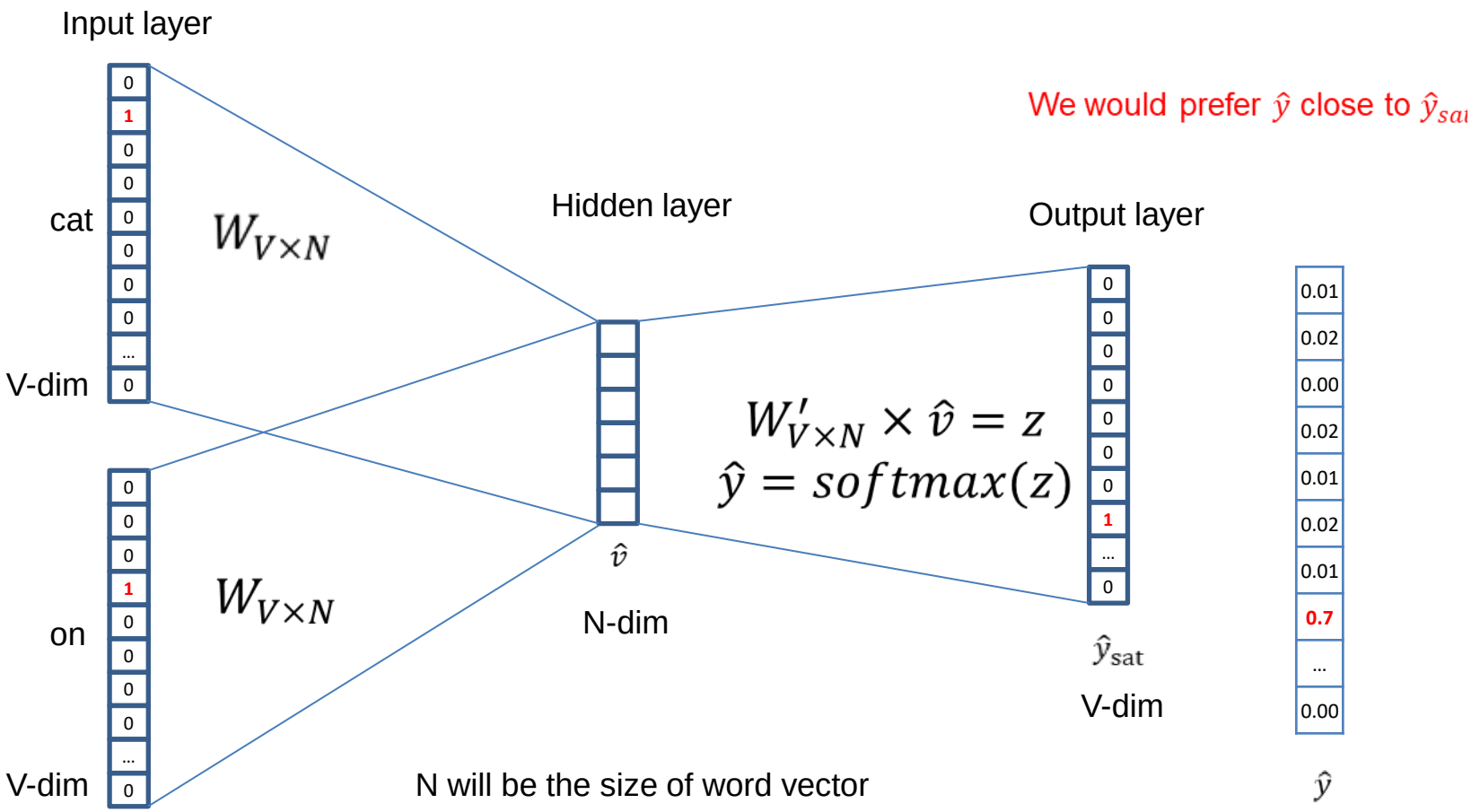
$$f(x) = \frac{1}{1 + e^{-x}}$$

- Softmax – обобщение применения логистической функции для многомерного случая

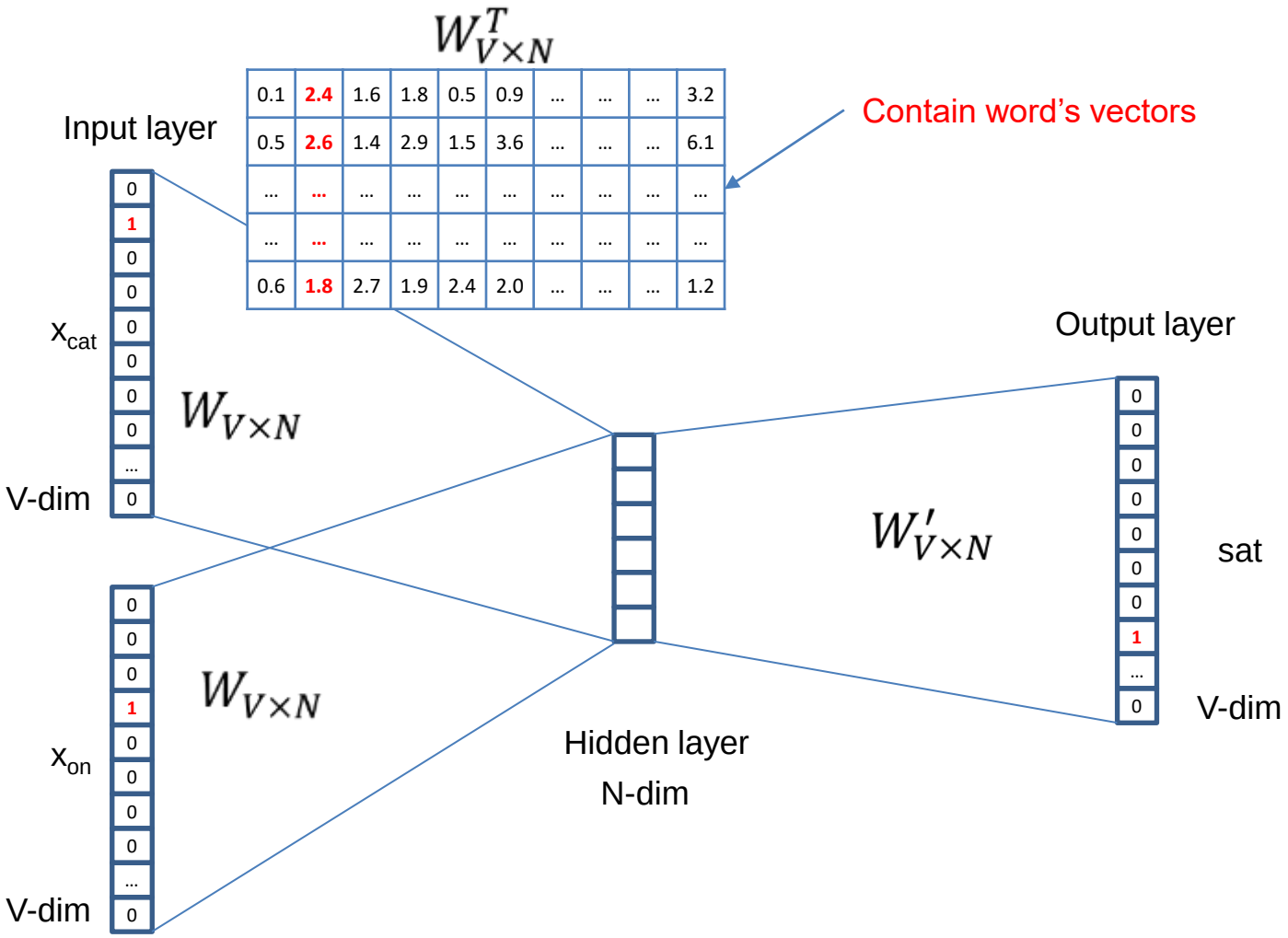
$$p_i = \mathrm{softmax}(s_i, \vec{s}) = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

- Softmax повышает максимальную величину и «прижимает» меньшие величины
- Примеры: сеть предсказала значения
- [1,2,3,**4**,1,2,3] -> softmax [0.024, 0.064, 0.175, **0.475**, 0.024, 0.064, 0.175]

Word2Vec: Continuous bag of words



Word2Vec: Continuous bag of words



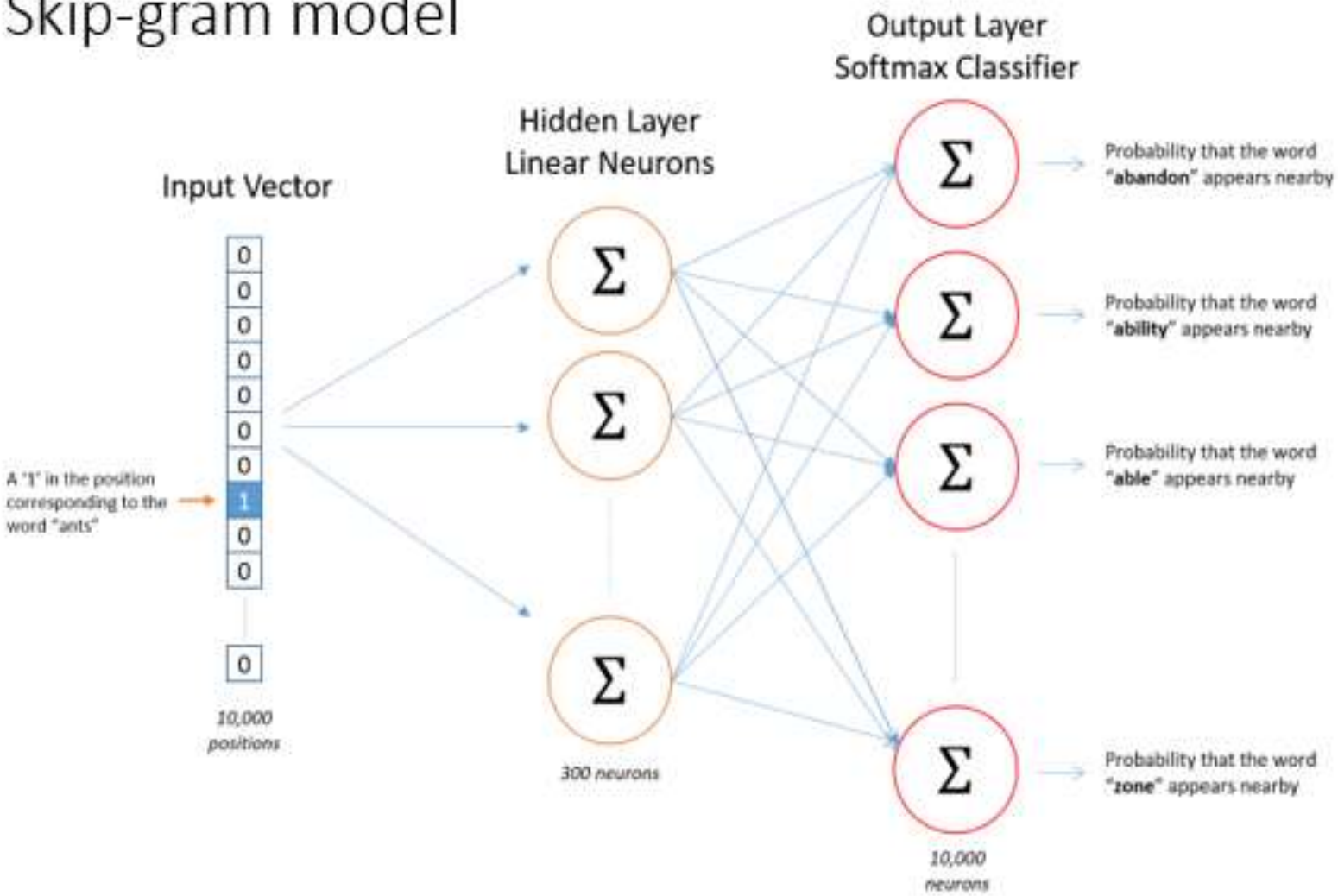
И W, и W' (представление слов контекста) можно рассматривать как представления слов. Но word2vec - W

Вектора word2vec

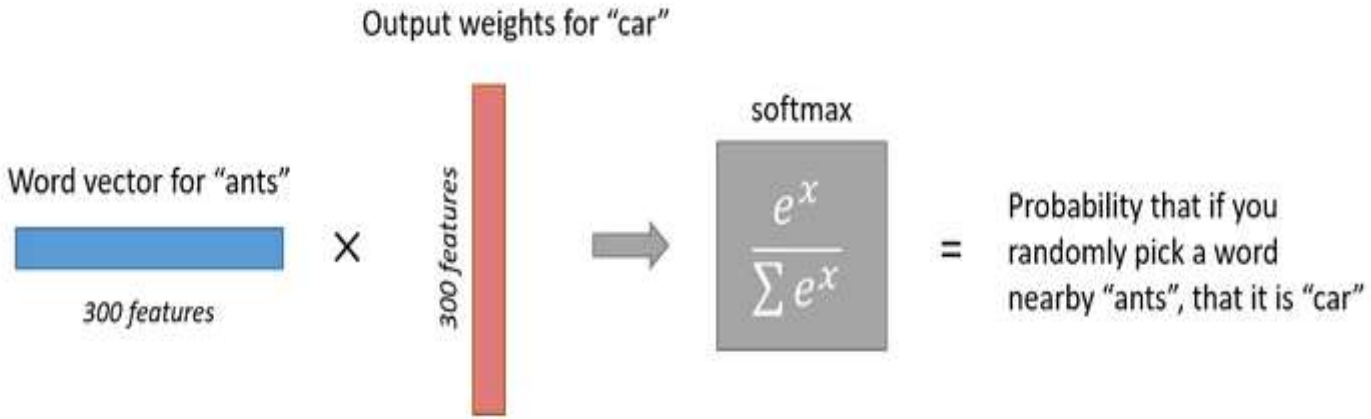
- Для каждого слова порождаются два вектора
  - Вектор слова как целевого
  - Вектор слова как контекст
  - Это разные вектора
  - Фактически мы хотим, чтобы вектор целевого слова был похож на вектора слов в контексте.

Word2vec

Skip-gram model



Предсказание слова car около слова ants



Большие матрицы, долгая обработка

- Подходы
  - Subsampling frequent words
    - Для the – слишком много контекстов,
    - The – мало что говорит о соседних словах
    - <u>Решение</u>
      - Слова выкидываются из текста с вероятностью пропорциональной их частоте
  - Negative sampling (негативное сэмплирование)

Негативное сэмплирование

- Обучение по каждому примеру требует пересчета весов для всех слов в выходном слове
- Идея: выбрать некоторое количество (например, 5) негативных слов (т.е. тех которых нет в контексте) и только для них перестроить веса (и также веса пересчитываются для положительных слов)
- Негативные слова выбираются с вероятностью, связанной с их частотой. В реализации word2vec – это выглядит так:

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^{n} \left(f(w_j)^{3/4} \right)}$$

Реализации word2vec

- Исходный код:
  - https://github.com/tmikolov/word2vec
- Gensim
  - https://radimrehurek.com/gensim/models/word2vec.html
- Есть реализации в пакетах нейронных сетей (Torch, TensorFlow, Theano)
- Есть уже насчитанные модели
  - Для русского языка (rusvectores.org)

Подход FastText

- В качестве векторного представления для слова берется среднее из представлений входящих в него n-грамм
  - то есть слово «ребенок» - это некоторая усредненная сумма векторов «ре», «еб», «бе», «ен» и т.д. (пример для биграмм)
- Плюсы подхода:
  - Символьные n-граммы встречаются чаще, чем слова целиком.
  - Учтем похожесть контекстов слов с одинаковыми аффиксами.
    - n-граммы из аффиксов «выловят» семантику и синтаксис,
    - n-граммы из корней – лексику.

Модели

В настоящий момент вы можете скачать следующие модели (жирным выделены модели, доступные для выбора в веб-интерфейсе):

Идентификатор	Скачать	Корпус	Размер корпуса	Объём словаря	Частотный порог	Тагсет	Ал
ruscorpora_upos_skipgram_300_5_2018	191 Мбайт	НКРЯ	250 миллионов слов	195 071	20	Universal Tags	Co Sk
ruwikiruscorpora_upos_skipgram_300_2_2018	376 Мбайт	НКРЯ и Википедия за декабрь 2017	600 миллионов слов	384 764	40	Universal Tags	Co Sk
news_upos_cbow_600_2_2018	547 Мбайт	Русскоязычные новости,с сентября 2013 до ноября 2016	почти 5 миллиардов слов	289 191	200	Universal Tags	Co Ba W
araneum_upos_skipgram_300_2_2018	192 Мбайта	Araneum Russicum Maximum	около 10 миллиардов слов	196 620	400	Universal Tags	Co Sk
araneum_none_fasttextcbow_300_5_2018	1 Гбайт	Araneum Russicum Maximum	около 10 миллиардов слов	195 782	400	Нет	fas CE (3. гра
araneum_none_fasttextskipgram_300_5_2018	675 Мбайт	Araneum Russicum Maximum	около 10 миллиардов слов	195 782	400	Нет	fas Sk (3- гра
ruwikiruscorpora-nobigrams_upos_skipgram_300_5_2018	385 Мбайт	НКРЯ и Википедия за декабрь 2017 (без склеивания биграмм)	600 миллионов слов	394 332	40	Universal Tags	Co Sk

Сайт rusvectores.ru: сходство векторов слов

Semantic associates for *стол* (ALL)

Ruscorpora and Russian Wikipedia

1. столик 0.679
2. табурет 0.526
3. табуретка 0.515
4. подоконник 0.501
5. диван 0.491
6. стул 0.484
7. кровать 0.476
8. тумбочка 0.447
9. парта 0.439
10. кушетка 0.428

Ruscorpora

1. столик 0.794
2. подоконник 0.642
3. табуретка 0.637
4. табурет 0.623
5. диван 0.582
6. кровать 0.573
7. стул 0.570
8. кушетка 0.561
9. тумбочка 0.561
10. кресло 0.552

Web corpus

1. столик 0.637
2. стул 0.570
3. табурет 0.554
4. поднос 0.525
5. тумбочка 0.517
6. табуретка 0.497
7. обеденный 0.490
8. кушетка 0.482
9. кресло 0.479
10. сервировочный 0.470

Тестирование моделей

- Датасеты пар слов с человеческими оценками семантической близости
  - Семантическое сходство: собака – животное, можно заменить одно слово другим,
  - Семантическая ассоциация: муж – жена, собака – конура, машина – бензин

- Вставка синонимов для слова в контексте
- Пропорции-аналогии

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
  - Rubenstein & Goodenough

- 3.92 car automobile
- 3.84 gem jewel
- 1.66 lad brother
- 1.68 crane implement
- 0.08 rooster voyage
- 0.08 noon string
  - Miller & Charles

Wordsim similarity – 353 пары слов

Семантическая близость

tiger	cat	8.31
tiger	tiger	10.00
plane	car	5.85
train	car	6.31
television	radio	6.85
media	radio	8.08
bread	butter	7.54
cucumber	potato	5.92

- Семантическая ассоциация

– book	paper	7.69
– computer	keyboard	8.38
– computer	internet	8.23
– telephone	communication	8.38
– drug	abuse	6.46
– smart	student	6.58
– company	stock	6.85
– stock	market	7.46
– stock	phone	1.54

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 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.

Результаты WordSim-353. Ранние подходы.
Корреляция Спирмена – чем выше, тем лучше

L&C	Leacock and Chodorow (1998)	Hassan and Mihalcea (2011)	Knowledge-based	0.302	0.356
WNE	Jarmasz (2003)	Hassan and Mihalcea (2011)	Knowledge-based	0.305	0.271
J&C	Jiang and Conrath 1997	Hassan and Mihalcea (2011)	Knowledge-based	0.318	0.354
L&C	Leacock and Chodorow (1998)	Hassan and Mihalcea (2011)	Knowledge-based	0.348	0.341
H&S	Hirst and St-Onge (1998)	Hassan and Mihalcea (2011)	Knowledge-based	0.302	0.356
Lin	Lin (1998)	Hassan and Mihalcea (2011)	Corpus-based	0.348	0.357
Resnik	Resnik (1995)	Hassan and Mihalcea (2011)	Knowledge-based	0.353	0.365
ROGET	Jarmasz (2003)	Hassan and Mihalcea (2011)	Knowledge-based	0.415	0.536

Результаты WordSim-353. Недавние подходы

DSG	Song et al. (2018)	Song et al. (2018)	Corpus-based	0.726	N/A
Sa18	Salle et al. (2018)	Dobó (2019)	Corpus-based	0.733	0.704
ESA	Gabrilovich and Markovitch (2007)	Gabrilovich and Markovitch (2007)	Corpus-based	0.748	0.503
Do19-hybrid	Dobó (2019)	Dobó (2019)	Hybrid	0.795	0.276
TSA	Radinsky et al. (2011)	Radinsky et al. (2011)	Hybrid	0.80	N/A
CLEAR	Halawi et al. (2012)	Halawi et al. (2012)	Corpus-based	0.81	N/A
Y&Q	Yih and Qazvinian (2012)	Yih and Qazvinian (2012)	Hybrid	0.81	N/A
ConceptNet Numberbatch	Speer et al. (2017)	Speer et al. (2017)	Hybrid	0.828	N/

Сбор данных о семантической близости для русских слов

№	Первое слово	Второе слово	Насколько *связаны* эти слова?
1	чашка	напиток	не связаны
2	автомобиль	машина	не связаны
3	производство	поход	не связаны
4	Арафат	терроризм	не связаны
5	поездка	машина	не связаны
6	психология	наука	не связаны
7	перечисление	категория	не связаны
8	среда	новости	не связаны
9	любовь	секс	не связаны

Example of human judgements about semantic similarity (HJ)

word1	word2	sim
петух (cock)	петушок (cockerel)	0.952
побережье (coast)	берег (shore)	0.905
тип (type)	вид (kind)	0.852
миля (mile)	километр (kilometre)	0.792
чашка (cup)	посуда (tableware)	0.762
птица (bird)	петух (cock)	0.714
война (war)	войска (troops)	0.667
улица (street)	квартал (block)	0.667
...	...	...
доброволец (volunteer)	девиз (motto)	0.091
аккорд (chord)	улыбка (smile)	0.088
энергия (energy)	кризис (crisis)	0.083
бедствие (disaster)	площадь (area)	0.048
производство (production)	экипаж (crew)	0.048
мальчик (boy)	мудрец (sage)	0.042
прибыль (profit)	предупреждение (warning)	0.042
напиток (drink)	машина (car)	0.000
сахар (sugar)	подход (approach)	0.000
лес (forest)	погост (graveyard)	0.000
практика (practice)	учреждение (institution)	0.000

Тестирование: тест на синонимы

ESL Synonym Tests

- Provide one target word in context
- Select “closest” synonym from a list of 4
- Used in previous versions of TOEFL and other standardized tests

- http://aclweb.org/aclwiki/index.php?title=ESL_Synonym_Questions_(State_of_the_art)
- 50 question data set available from Peter Turney
  - http://www.apperceptual.com/

Пример теста на синонимы

- Stem: "A *rusty* nail is not as strong as a clean, new one."
- Choices:
  - (a) corroded
  - (b) black
  - (c) dirty
  - (d) painted

Датасеты для тестирования по аналогиии

- Аналогии “a is to a ∗ as b is to b∗
  - MSR’s analogy dataset
    - 8000 морфосинтаксических аналогий
    - good is to best as smart is to smartest
  - Google’s analogy dataset
    - 19544 морфосинтаксических и семантических аналогий
    - Paris is to France as Tokyo is to Japan

Word Analogies

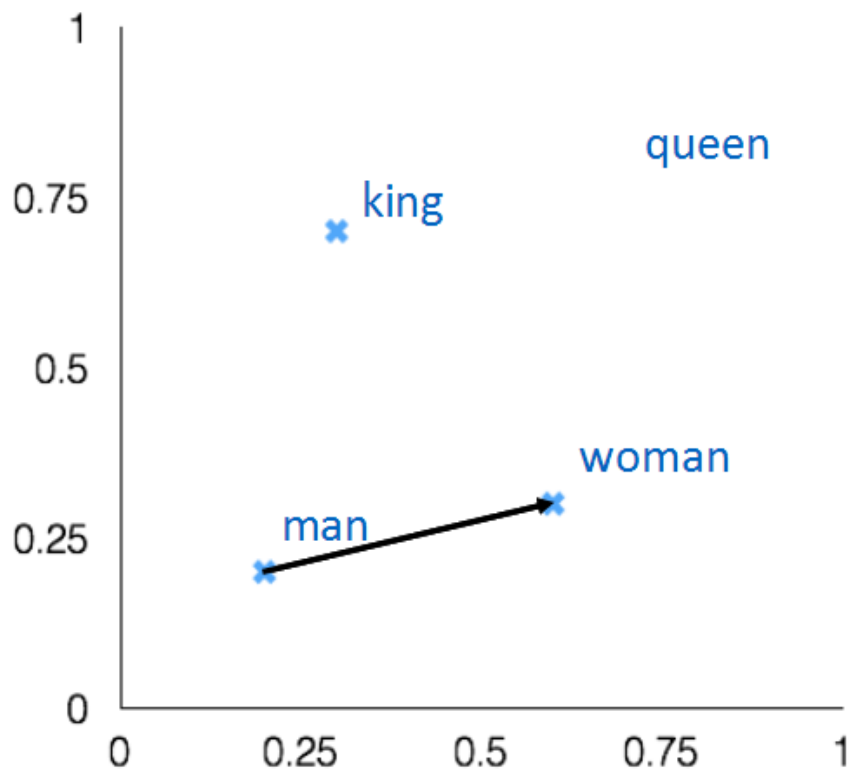
Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:?

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

king	[0.30 0.70]
man	[0.20 0.20]
woman	[0.60 0.30]
queen	[0.70 0.80]



RusVectōrēs

Похожие слова | Визуализации | Калькулятор | Различные операции | Модели | О проекте | Контакты | RU/EN

Семантический калькулятор

Вы можете вычислять отношения. Например, **«найти слово D, связанное со словом C таким же образом, как слово A связано со словом B»**. Таким образом можно определять семантические связи между понятиями. В форме ввода приведен пример: какое слово относится к слову **«лондон»**, так же, как **«россия»** относится к **«москве»**? Ответ — **«великобритания»**: Лондон столица Великобритании, а Москва — столица России. Подробнее...

москва_PROPN → россия_PROPN

варшава_PROPN → ???

Новостной корпус

1. польша 0.41
2. белоруссия 0.39
3. страна 0.39
4. германия 0.38
5. европа 0.38

Выберите модель:

☑ Новостной корпус ☐ Araneum fastText ☐ НКРЯ и Wikipedia ☐ НКРЯ

Показывать только:

○ Прилагательные ○ Имена собственные ○ Глаголы ○ Существительные ○ Наречия
◉ Все части речи

Вычислить!

Вы также можете попробовать более сложные операции над векторами, чем простое решение пропорции.

Введите в **«положительную»** и **«отрицательную»** формы не более 10 слов через пробел. *RusVectōrēs* сложит вектора положительных слов и вычтет из них отрицательные. Затем он выдаст слова, наиболее близкие к получившемуся вектору. Если вы оставите отрицательное поле пустым, *RusVectōrēs* просто найдет центр лексического кластера, образованного положительными словами.

+

телефон маленький

-

Выберите модель:

☑ Новостной корпус ☐ Araneum fastText ☐ НКРЯ и Wikipedia
☐ НКРЯ

Показывать только:

○ Прилагательные ○ Имена собственные ○ Глаголы
○ Существительные ○ Наречия ◉ Все части речи

Вычислить!

Сексизм нейронных языковых моделей

RusVectōrēs

Похожие слова | Визуализации | Калькулятор | Различные операции | Модели | О проекте | Контакты | RU/EN

Семантический калькулятор

Вы можете вычислять отношения. Например, **«найти слово D, связанное со словом C таким же образом, как слово A связано со словом B»**. Таким образом можно определять семантические связи между понятиями. В форме ввода приведен пример: какое слово относится к слову **«лондон»**, так же, как **«россия»** относится к **«москве»**? Ответ — **«великобритания»**: Лондон столица Великобритании, а Москва — столица России. Подробнее...

мужчина_NOUN → программист_NOU

женщина_NOUN → ???

Новостной корпус

1. швея 0.43
2. логист 0.41
3. компьютерщик 0.40
4. тестировщик 0.39
5. продажник 0.39

Araneum fastText

1. программистка 0.83
2. програмист 0.82
3. программер 0.74
4. программинг 0.72
5. программистский 0.67

Выберите модель:

☑ Новостной корпус ☑ Araneum fastText ☐ НКРЯ и Wikipedia ☐ НКРЯ

Показывать только:

○ Прилагательные ○ Имена собственные ○ Глаголы ○ Существительные ○ Наречия ◉ Все части речи

Вычислить!

Вы также можете попробовать более сложные операции над векторами, чем простое решение пропорции.

Введите в **«положительную»** и **«отрицательную»** формы не более 10 слов через пробел. *RusVectōrēs* сложит вектора положительных слов и вычтет из них отрицательные. Затем он выдаст слова, наиболее близкие к получившемуся вектору. Если вы оставите отрицательное поле пустым, *RusVectōrēs* просто найдет центр лексического кластера, образованного положительными словами.

\+

телефон маленький

\-

Выберите модель:

☑ Новостной корпус ☑ Araneum fastText ☐ НКРЯ и Wikipedia ☐ НКРЯ

Показывать только:

○ Прилагательные ○ Имена собственные ○ Глаголы ○ Существительные ○ Наречия ◉ Все части речи

Вычислить!

Слова, выделенные зеленым, являются высокочастотными (доля слова в корпусе выше 0.00001); слова, выделенные красным, являются низкочастотными (доля слова в корпусе ниже 0.0000005).

Применение векторных представлений слов

- 1) Исследование семантических близостей слов
  - Группирование слов,
  - Предсказание отношений между словами

- 2) Вход всех нейронных сетей для автоматической обработки текстов – это векторные представления слов (+ доп. признаки()
  - Эффективно брать предобученные векторные представления

Задание. Wordsim similarity

- Взять датасеты
  - Wordsim-353 http://alfonseca.org/pubs/ws353simrel.tar.gz
  - Similarity: **wordsim_similarity_goldstandard.txt**:
  - Relatedness: **wordsim_relatedness_goldstandard.txt**
  - Simlex-999 https://fh295.github.io/simlex.html

- Модель Fasttext https://fasttext.cc/docs/en/pretrained-vectors.html.
- Посчитать близости для пар слов из списков wordsim similarity, wordsim-relatedness.
- Вычислить корреляцию Спирмена (пакет) результатов с ручными оценками
- Есть ли корреляция модели с человеческими оценками?
- Для какого списка корреляция выше? Почему результаты могут так различаться