Векторная семантика

(на основе материалов Jurafsky & Martin)

Семантика (определение в общем виде)

Семантика — раздел лингвистики, который посвящен изучению смыслов слов и смысловых отношения между словами.

Семантика — синоним слова “смысл”.

Значения слов

Как узнать значение слова?

Например, посмотреть в словаре.

Слово “кролик”

Кролик (значения из Wiki-словаря)

1. зоол. небольшое пушистое животное с длинными ушами, млекопитающее из семейства зайцев ◆ Имеется кроличатник, где живет до двух с половиной тысяч **кроликов**, по преимуществу породы «Великан Белый» и «Фландры». *Н. Х. Виленская, В. Н. Клычин, «На лыжах по окрестностям Ленинграда», 1930 г. (цитата из Национального корпуса русского языка, см. Список литературы)*
2. разг. мех этого животного ◆ Шапка из **кролика**.
3. мясо этого животного, употребляемое в пищу ◆ Могу вполне прилично приготовить какое-нибудь традиционное французское блюдо, например, **кролика**. *Ирина Милеева, «Какая встреча! Кристиан Лиэгр», 2004 г. // «Homes & Gardens» (цитата из Национального корпуса русского языка, см. Список литературы)*

Отношения между словами

Антонимы

тепло / холод

сухой / мокрый

добрый / злой

старье / новинка

падать / подниматься

вниз / вверх

Отношения между словами

Синонимы (имеют почти (!) одинаковое значение)

большущий / огромный

вода / H_2O

смелый / храбрый

малыш / карапуз

Отношения между словами

1) **Слова со схожим компонентом в значении (англ. similar)**
   машина, велосипед
   корова, лошадь

2) **Слова из одной области (англ. related)**
   машина, велосипед
   машина, бензин

Отношения между словами

Иерархические отношения

Итак, семантика

1) У слов бывает несколько значений
2) Типы отношений между значениями слов
   - ➢ антонимия
   - ➢ синонимия
   - ➢ сходство (relatedness)
   - ➢ гипонимия/гиперонимия

Основная идея “векторной семантики”

Firth (1957): *“You shall know a word by the company it keeps!”*

- A bottle of **tesgüino** is on the table
- Everybody likes **tesgüino**
- **Tesgüino** makes you drunk
- We make **tesgüino** out of corn.

Что такое tesgüino?

Основная идея “векторной семантики”

Firth (1957): *“You shall know a word by the company it keeps!”*

- ➢ A bottle of **tesgüino** is on the table
- ➢ Everybody likes **tesgüino**
- ➢ **Tesgüino** makes you drunk
- ➢ We make **tesgüino** out of corn.

Что такое tesgüino?

По предложениям можно понять, что tesgüino — это напиток вроде пива.

- **Два слова имеют схожее значение, если они встречаются в похожих контекстах. Слова можно перевести в векторное представление, посчитав их встречаемость с другими словами.**

Основная идея “векторной семантики”

Все это прекрасно, но ведь у нас есть словари и тезаурусы, зачем все усложнять?

- ➢ У нас нет словарей и тезаурусов для всех языков.

Основная идея “векторной семантики”

Все это прекрасно, но ведь у нас есть словари и тезаурусы, зачем все усложнять?

- ➢ У нас нет словарей и тезаурусов для всех языков.
- ➢ Словари и тезаурусы обновляются не очень часто, а значит, нельзя проследить изменения значений слов во времени.

Основная идея “векторной семантики”

Все это прекрасно, но ведь у нас есть словари и тезаурусы, зачем все усложнять?

- ➢ У нас нет словарей и тезаурусов для всех языков.
- ➢ Словари и тезаурусы обновляются не очень часто, а значит, нельзя проследить изменения значений слов во времени.
- ➢ Ограниченность наполнения, не все слова есть в словаре.

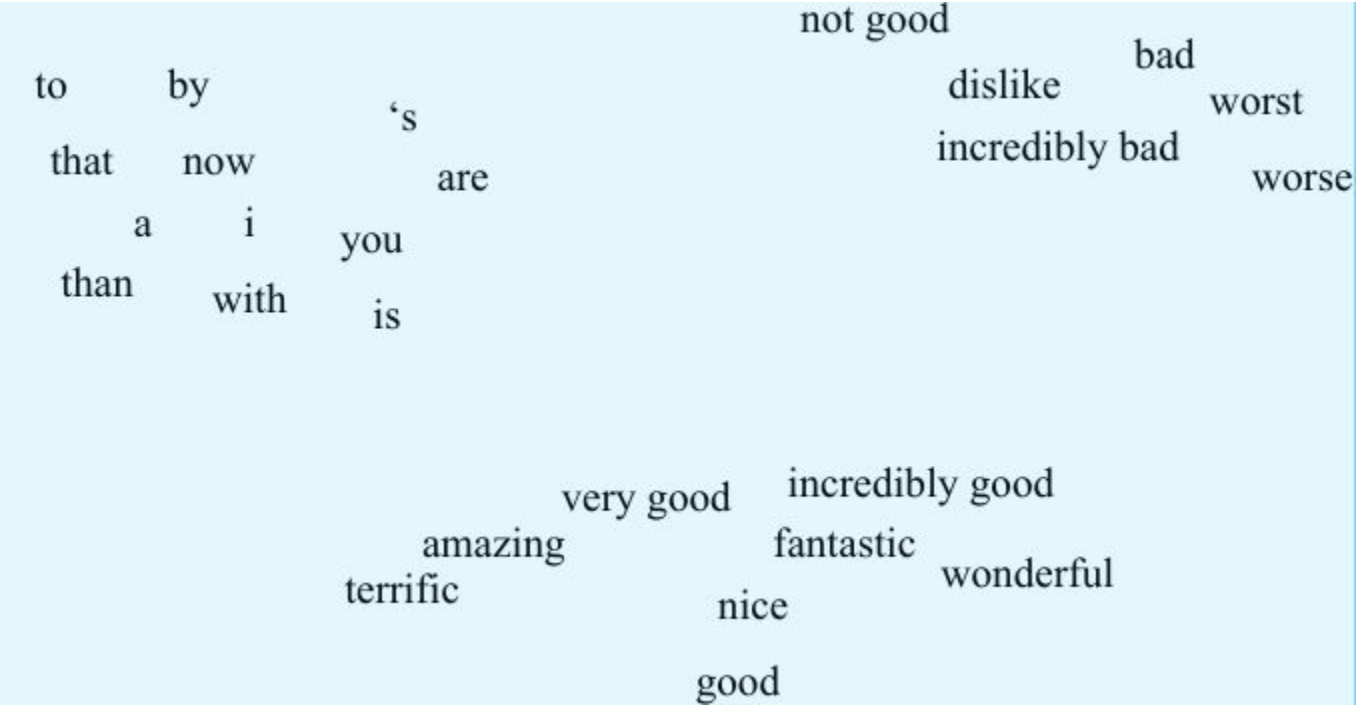
Основная идея “векторной семантики”

Все это прекрасно, но ведь у нас есть словари и тезаурусы, зачем все усложнять?

- ➢ У нас нет словарей и тезаурусов для всех языков.
- ➢ Словари и тезаурусы обновляются не очень часто, а значит, нельзя проследить изменения значений слов во времени.
- ➢ Ограниченность наполнения, не все слова есть в словаре.
- ➢ Проблемы с описанием глаголов и прилагательных.

Вектора слов

Похожие слова расположены рядом в пространстве



Вектора слов, word vectors, embeddings

- Смыслы — это вектора в многомерном пространстве или эмбеддинги.

Вектора слов, word vectors, embeddings

- Смыслы — это вектора в многомерном пространстве или эмбеддинги.
- *Called an "embedding" because it's embedded into a space.*

Вектора слов, word vectors, embeddings

- Смыслы — это вектора в многомерном пространстве или эмбеддинги.
- *Called an "embedding" because it's embedded into a space.*
- Слово представлено в виде вектора значений.

Вектора слов, word vectors, embeddings

- Смыслы — это вектора в многомерном пространстве или эмбеддинги.
- *Called an "embedding" because it's embedded into a space.*
- Слово представлено в виде вектора значений.
- В области NLP это стандартный способ представления значений слов и отношений между ними. Используется для решения множества задач, в том числе sentiment анализ, определение плагиата, вопросно-ответные системы и т.д.

Пример: два типа векторного представления

Tf-Idf

Baseline подход

Очень большое число измерений (разреженные матрицы)

Значения для слов считаются по простой формуле

Word2vec

Вектора меньшей размерности

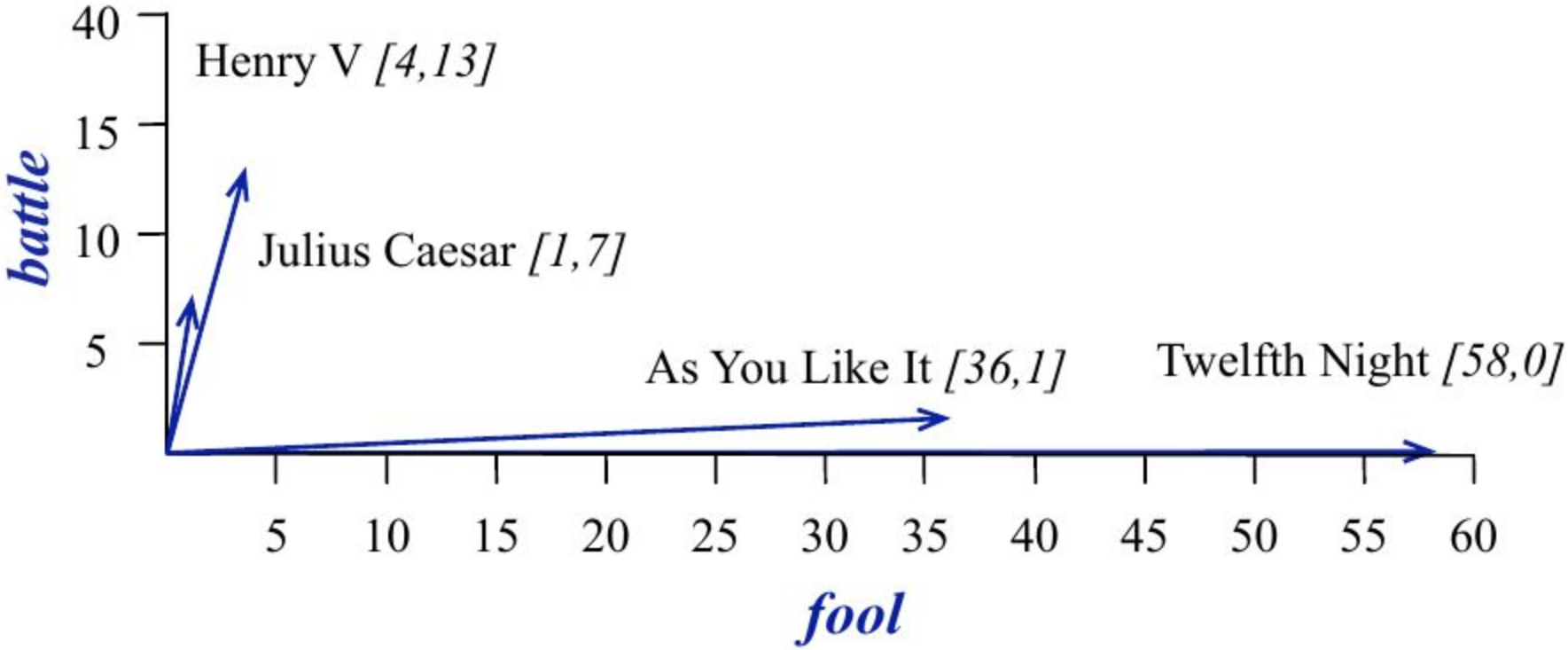
Значения для слов получаются в результате обучения модели

Матрица слово-документ

Каждый документ представлен в виде частот слов (Bag-of-words) и является вектором.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Вектора в пространстве



Извлечение информации

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Вектора комедий похожи и отличаются от трагедий.

В комедиях больше про дураков и шутки, нежели про сражения.

Вектора слов

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

battle — это такое слово, которое встречается в Julius Caesar и Henry V

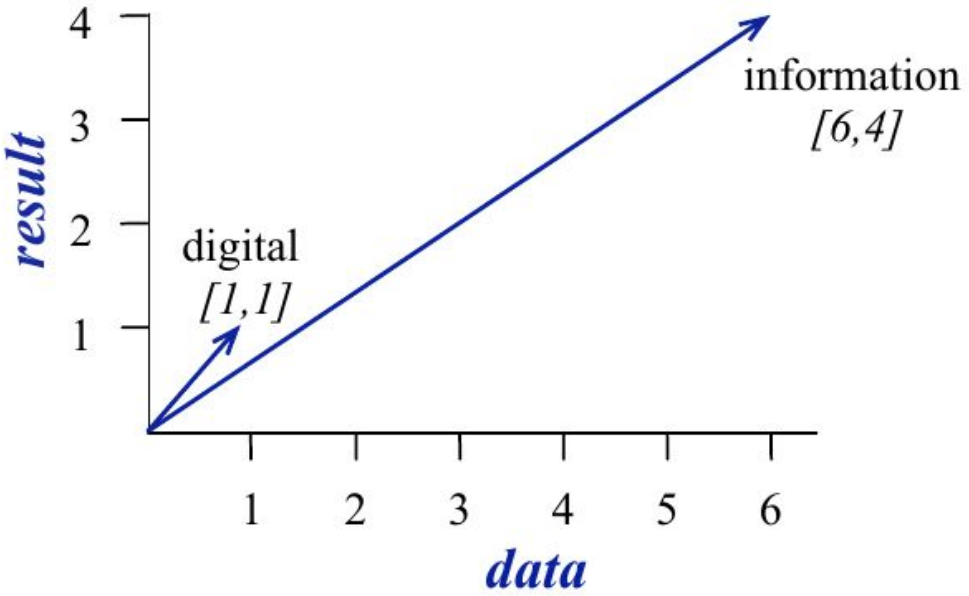
fool — это такое слово, которое встречается в комедиях и в особенности в Twelfth Night.

Матрица слово-слово

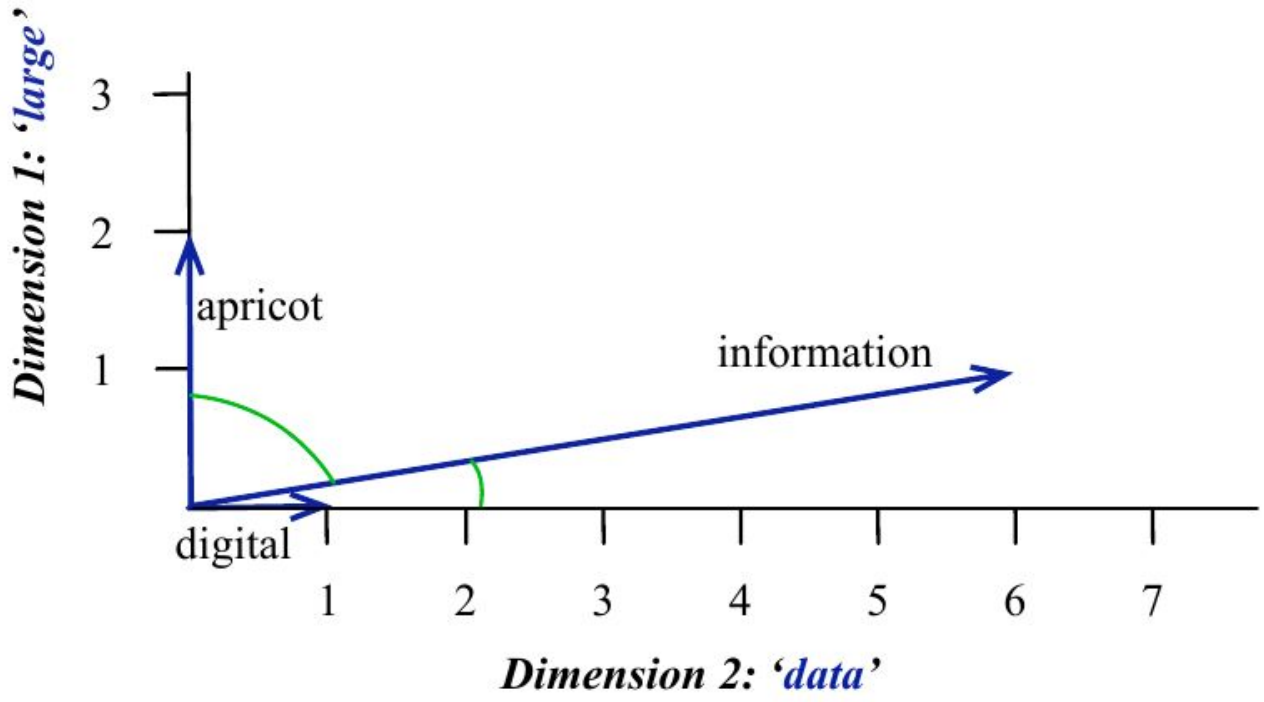
Два слова имеют схожее значение, если встречаются в похожих контекстах.

	aardvark	computer	data	pinch	result	sugar	...
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

Вектора слов в пространстве



Вектора слов в пространстве



Косинусная мера (cosine similarity)

Как интерпретировать?

-1: вектора направлены в противоположные стороны, значения слов сильно отличаются друг от друга

+1: вектора направлены в одну сторону, значения слов похожи

Косинусная мера (cosine similarity)

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

	large	data	computer
apricot	1	0	0
digital	0	1	2
information	1	6	1

Which pair of words is more similar?

cosine(apricot,information) =

$$\frac{1+0+0}{\sqrt{1+0+0}\ \sqrt{1+36+1}} = \frac{1}{\sqrt{38}} = .16$$

cosine(digital,information) =

$$\frac{0+6+2}{\sqrt{0+1+4}\ \sqrt{1+36+1}} = \frac{8}{\sqrt{38}\sqrt{5}} = .58$$

cosine(apricot,digital) =

$$\frac{0+0+0}{\sqrt{1+0+0}\ \sqrt{0+1+4}} = 0$$

Матрицы слово-слово

- Вместо документов берутся меньшие части текстов: отрывки, контекстное окно (± n слов).
- Слово представляется как подсчет совместной встречаемости с другими словами в контексте.
- Размерность вектора определяется количеством уникальных слов.
- Матрица “квадратная” V x V, где V — количество уникальных слов.

Особенности матриц слово-слово

- ➢ Слишком большая размерность. На примере 4x6, в действительности ~50000x50000
  - Матрица очень разреженная, много нулей
  - Это не является проблемой, т.к. существует много подходов к уменьшению размерности
- ➢ Размер окна контекста влияет на характер получаемой информации
  - Окно ± 1--3 позволяет зафиксировать скорее синтаксическую информацию.
  - Окно ± 4-10 дает больший охват семантики.

Embeddings, dense vectors

- Альтернатива традиционным подходам вроде обычных частот и TF-IDF.
- Меньше размерность, проще использовать в машинном обучении.
- Позволяют обобщить информацию в отличие от эксплицитно заданных частот.
- Позволяют лучше кодировать синонимию.
- Применение на практике показало лучшие результаты.

Векторные модели, которые можно скачать

- Word2vec (Mikolov et al.)
  https://code.google.com/archive/p/word2vec/
- Fasttext
  http://www.fasttext.cc/
- Glove (Pennington, Socher, Manning)
  http://nlp.stanford.edu/projects/glove/

Особенности

Сходство (similarity) векторов слов зависит от размера контекстного окна C.

C = ±2 ближайшие слова к слову Hogwarts:

- Sunnydale
- Evernight

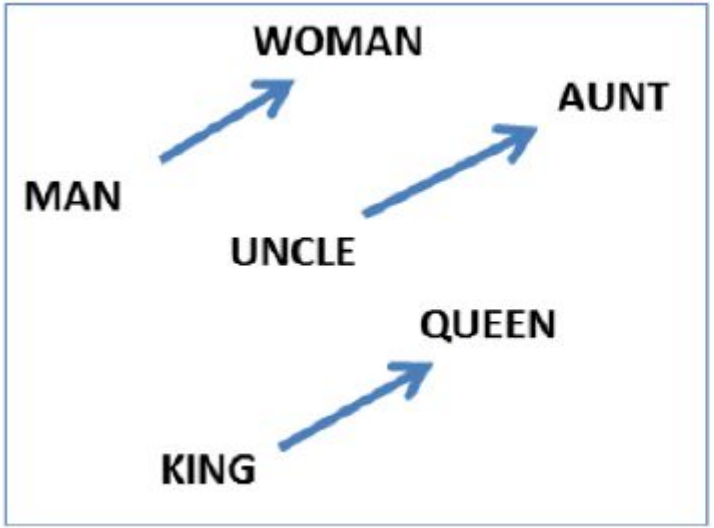
C = ±5 ближайшие слова к слову Hogwarts:

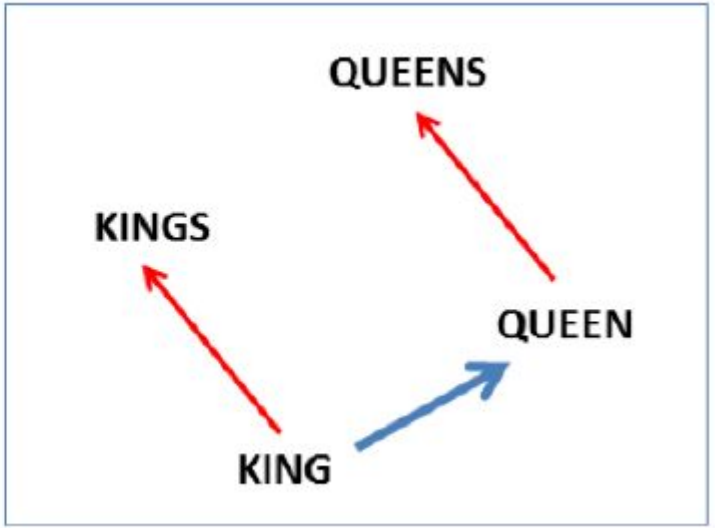
- Dumbledore
- Malfoy
- halfblood

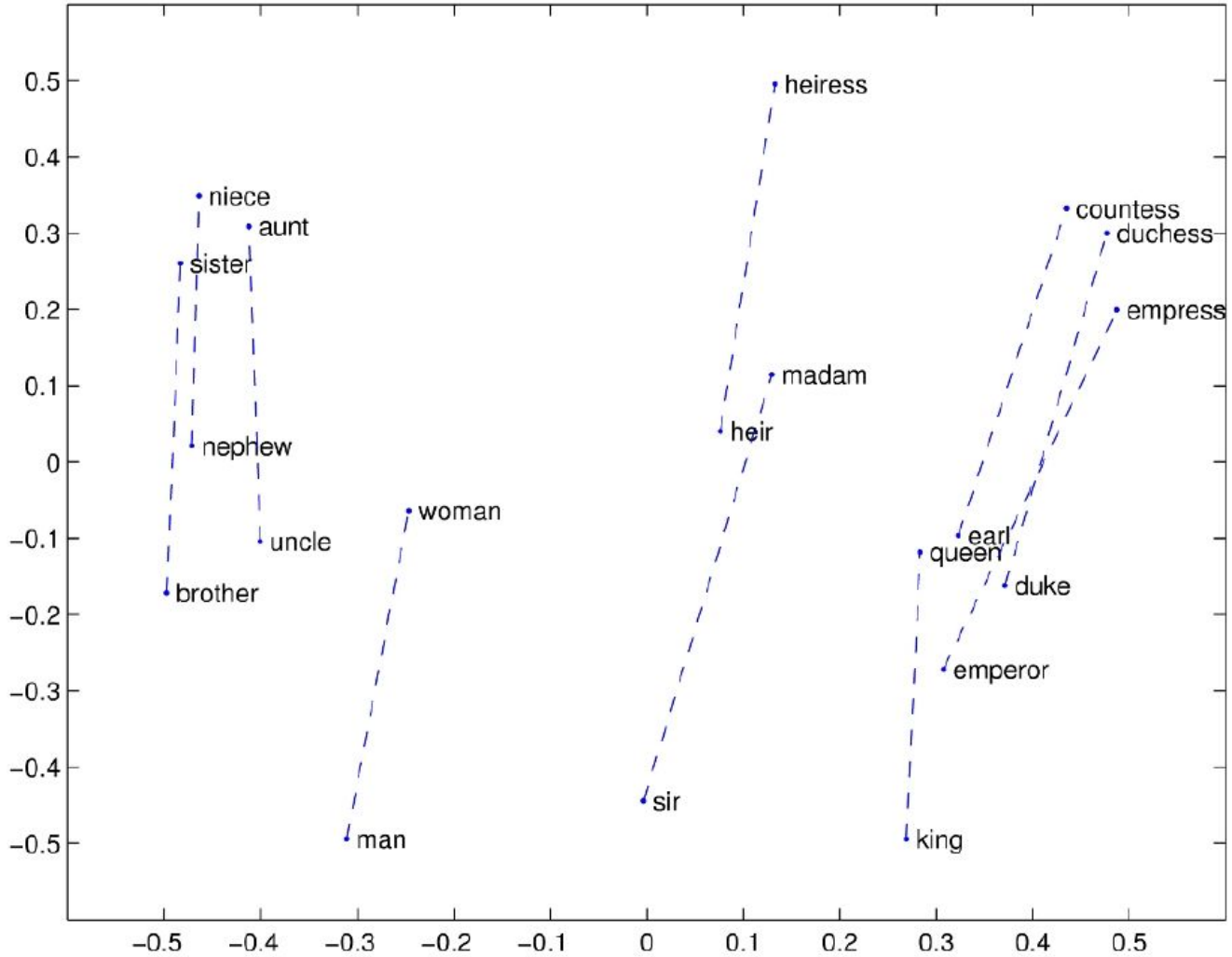
Пропорции (relational meaning)

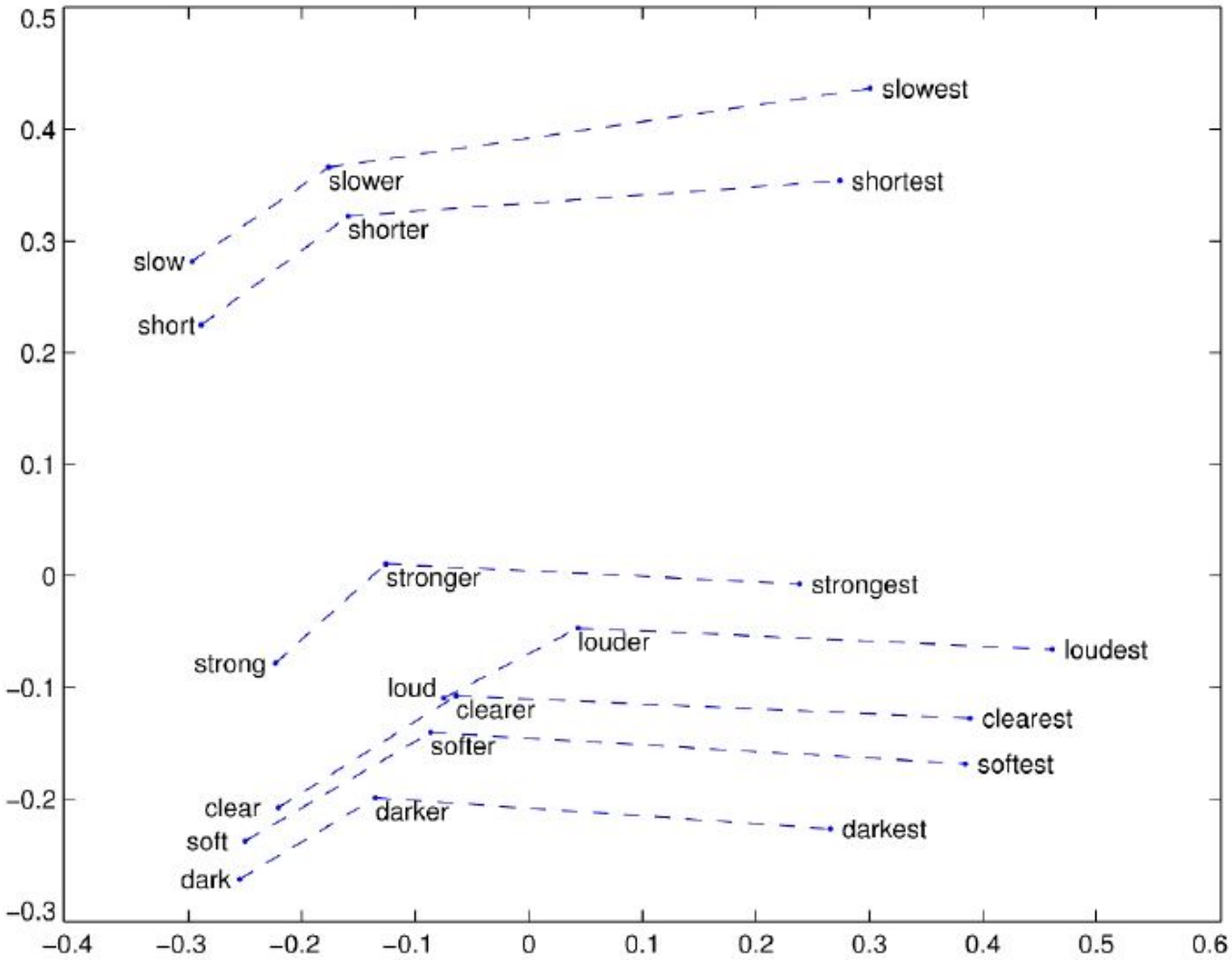
vector(‘king’) - vector(‘man’) + vector(‘woman’) ≈ vector(‘queen’)

vector(‘Paris’) - vector(‘France’) + vector(‘Italy’) ≈ vector(‘Rome’)






heiress
niece
aunt
sister
countess
duchess
empress
madam
heir
nephew
woman
uncle
earl
queen
duke
brother
emperor
sir
man
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6

Особенности

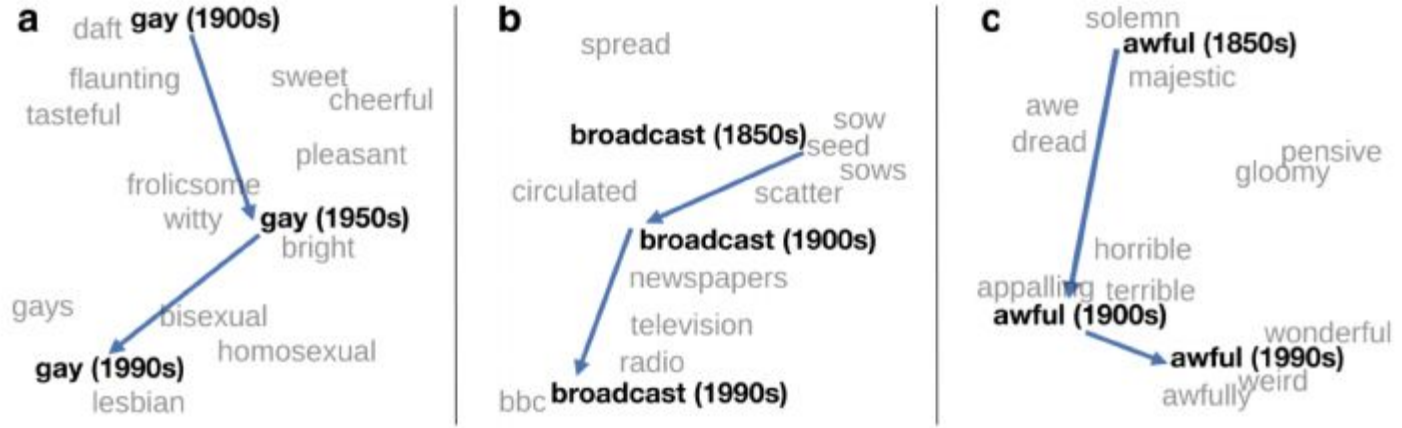
При помощи векторных моделей можно изучать, как изменялись значения слов в ходе истории.

Как?

Особенности

При помощи векторных моделей можно изучать, как изменялись значения слов в ходе истории.

Как? Обучить модель на текстах разного времени.



Культурные предубеждения (cultural bias)

- “Paris : France :: Tokyo : x”
  x = Japan

- “father : doctor :: mother : x”
  x = nurse

- “man : computer programmer :: woman : x”
  x = homemaker

Культурные предубеждения (cultural bias)

Caliskan, Aylin, Joanna J. Bruson and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356:6334, 183-186.

Результаты исследования:

- ➢ Вектора имен афро-американцев имели высокую близость с векторами слов *abuse, stink, ugly* в модели GloVe.
- ➢ Вектора европейских имен имели высокую близость с векторами слов *love, peace, miracle* в модели GloVe.
- ➢ Мужские имена ассоциировались у участников эксперимента с точными науками, женские — с искусством.

Модели фиксируют стереотипы, существующие в обществе.

Где и как работать с векторами слов?

- В Python есть библиотека gensim

  https://radimrehurek.com/gensim/models/word2vec.html

- RusVectōrēs: Семантические модели для русского языка.

  http://rusvectores.org/ru/