# ДИСТРИБУТИВНАЯ СЕМАНТИКА

Саша Ершова, asershova@edu.hse.ru

August 8, 2017

ЛОШ-17, Компьютерная лингвистика

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

Дистрибутивная семантика — способ применять математические модели к изучению значений слов. В частности, вычислять их ”похожесть”.

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

еда_NOUN

Найти похожие слова!

## Семантические аналоги для *еда*

1. пища 0.752
2. питие 0.637
3. жратва 0.612
4. готовка 0.600
5. поесть 0.597
6. ужин 0.556
7. всухомятку 0.552
8. лакомство 0.549
9. кормить 0.543
10. выпивка 0.543

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

Главное достоинство дистрибутивной семантики — мы можем применять её, даже не зная языка. Нужно только много текстов.

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

## Как это работает?

Дистрибутивная гипотеза: значение слова полностью определяется его контекстом. Похожие слова будут находиться в одних и тех же контекстах.

”You shall know a word by the company it keeps.” (Firth 1957)

# КАК ЭТО РАБОТАЕТ?

Например, слова ”еда” и ”пища” часто встречаются в похожих контекстах: ”очень вкусная еда/пища”, ”я приготовил еду/пищу” и т.д. Из этого мы можем сделать вывод, что слова ”еда” и ”пища” близки и часто взаимозаменяемы.

# КАК ЭТО РАБОТАЕТ?

1. Находим в корпусе все вхождения одного слова.
2. Для каждого вхождения берём контекст (например, 5 слов слева, 5 слов справа).
3. Обучаем нейросеть, которая на основе всех этих контекстов будет присваивать слову последовательность из N чисел. ”Вектор в N-мерном пространстве” — это просто упорядоченная последовательность из N чисел.
4. Повторяем для всех слов в корпусе.

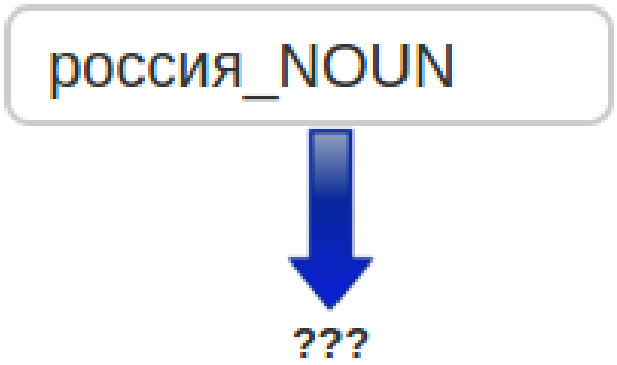
# КАК ЭТО РАБОТАЕТ?

При помощи нейросетей все слова в корпусе превращаются в векторы в 300-мерном (или другой размерности) пространстве, и потом можно смотреть на расположение векторов.

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

После того, как мы превратили все слова в векторы, можно что-нибудь считать семантическую близость слов и искать интересные пропорции.

# ДИСТРИБУТИВНАЯ СЕМАНТИКА


## НКРЯ и русская Wikipedia

1. пельмень 0.49
2. фаст фуд 0.44
3. сырник 0.43
4. шаурма 0.41
5. майонез 0.40

# ДИСТРИБУТИВНАЯ СЕМАНТИКА

Можно проводить математические операции над векторами (сложение, вычитание), и эти операции будут отражать реальную семантику слов.

Например:

- компьютер + маленький = ноутбук
- король - мужчина + женщина = королева
- человек - хороший = существо

Попробовать что-нибудь самостоятельно посчитать можно на http://rusvectores.org/ru/calculator/

# Спасибо за внимание!