Word embeddings

Елена Захарова

НИУ ВШЭ, 2020

Чем плох подход Bag-of-Words

- не учитывается взаимное расположение слов
- не учитывается семантика слова
- тяжело справляться с большими объемами данных (большой словарь + н-граммы => матрица признаков сильно раздувается и не влезает в память компьютера)

Гипотеза локальности

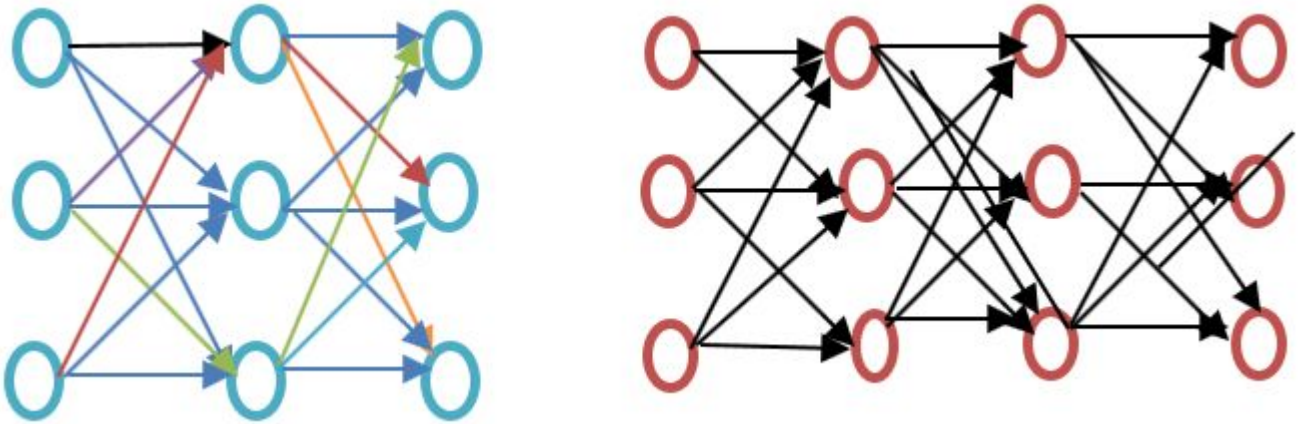
- Слова, встречающиеся в одинаковом окружении имеют похожие значения
- Проанализировав контексты, в которых встречается слово, можно что-то понять про его сочетаемость и про его семантику

Word2Vec

- Представить слова в виде векторов
- Векторы должны быть похожими у похожих слов
- Похожие слова = встречающиеся в схожих контекстах
- Используем для этого нейросеть
- Будем пытаться предсказать вероятность появления слова по его контексту (или наоборот)
- У похожих слов должны быть похожие вероятности

Word2Vec - что это

- Неглубокая нейросеть - 1 скрытый слой



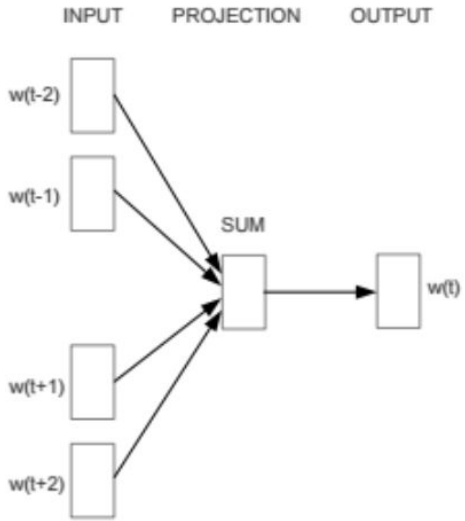
Shallow vs Deep Learning

Основная идея

- Ставим “искусственную” задачу предсказания вероятности
- Обучаем нейросеть на данных с этой задачей
- Берем вектор весов с внутреннего слоя, это и будет нужный нам вектор слова
- Векторы весов инициализируются рандомно, в процессе обучения меняются (векторы слов, появляющихся в одном контексте сближаются, векторы слов не встречающихся в одном контексте отдаляются)

Word2Vec CBoW vs Skip-gram

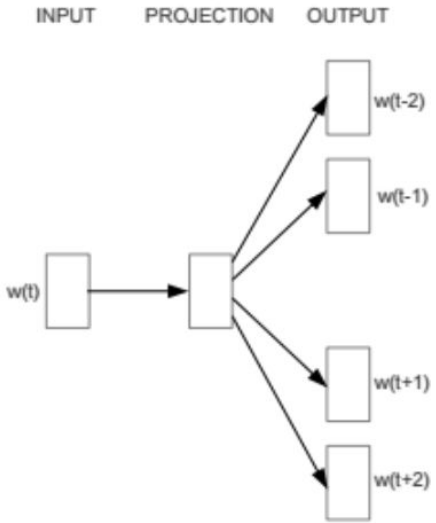
Given a set of (neighboring) words, **guess single words** that potentially occur along with this set of words.



CBOW

(Continuous Bag Of Words)

or



Skip-gram

Guess potential neighboring words based on the single word being analyzed.

Word2vec CBoW vs Skip-gram

- CBoW:
  - быстрее
  - лучше на частотных словах
- Skip-gram:
  - лучше работает на маленьком количестве данных
  - лучше улавливает слова с низкой частотностью