

word2vec

ПОХОЖИЕ СЛОВА

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

- › «Идти» и «шагать» — синонимы
- › Для компьютера это разные строки
- › Как понять, что они похожи?
- › На основе данных!
- › Слова со схожим смыслом часто идут в паре с одними и теми же словами
- › У них похожие контексты

- › Хотим каждое слово представить как вещественный вектор: $w \rightarrow \vec{w} \in \mathbb{R}^d$
- › Какие требования?
 - ▶ Размерность d должна быть не очень велика
 - ▶ Похожие слова должны иметь близкие векторы
 - ▶ Арифметические операции над векторами должны иметь смысл

- › Будем обучать представления слов так, чтобы они хорошо предсказывали свой контекст
- › Выборка состоит из текстов, каждый представляет собой последовательность слов $w_1, \dots, w_i, \dots, w_n$

$$\sum_{i=1}^n \sum_{j=-k}^k \log p(w_{i+j} | w_i) \rightarrow \max,$$

- › Выборка состоит из текстов, каждый представляет собой последовательность слов $w_1, \dots, w_i, \dots, w_n$

$$\sum_{i=1}^n \sum_{j=-k}^k \log p(w_{i+j}|w_i) \rightarrow \max,$$

где вероятность вычисляется через soft-max:

$$p(w_i|w_j) = \frac{\exp(\langle \vec{w}_i, \vec{w}_j \rangle)}{\sum_w \exp(\langle \vec{w}, \vec{w}_j \rangle)}$$

- › Косинусное расстояние хорошо отражает схожесть слов по тематике (в зависимости от корпуса)
- › $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$
- › $\vec{\text{Moscow}} - \vec{\text{Russia}} + \vec{\text{England}} \approx \vec{\text{London}}$
- › Перевод: $\vec{\text{oñe}} - \vec{\text{uño}} + \vec{\text{four}} \approx \vec{\text{quarto}}$
- › Среднее представление по всем словам в тексте — хорошее признаковое описание

word2vec И ОБУЧЕНИЕ С УЧИТЕЛЕМ

- › Проблема мешка слов — слишком большое количество признаков
- › Средний word2vec-вектор позволяет получить компактное признаковое описание
- › При размерности вектора 100 можно обучать композиции деревьев

- › word2vec позволяет описать каждое слово вектором
- › Похожие слова имеют близкие векторы
- › Признаки для текста — средний вектор по всем словам

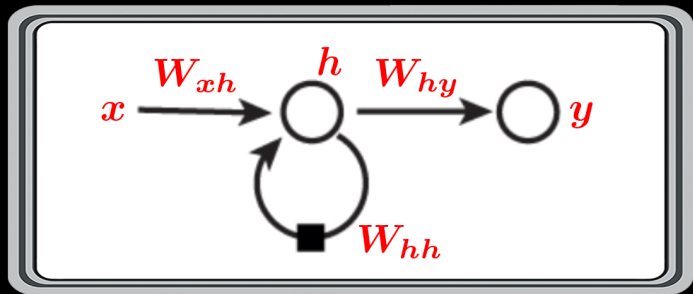
РЕКУРРЕНТНЫЕ СЕТИ

- › Считаем, что слова появляются в тексте независимо
- › n-граммы, skip-граммы — аналогично

- › Обучается с учётом контекста слов
- › При применении порядок слов всё равно не учитывается

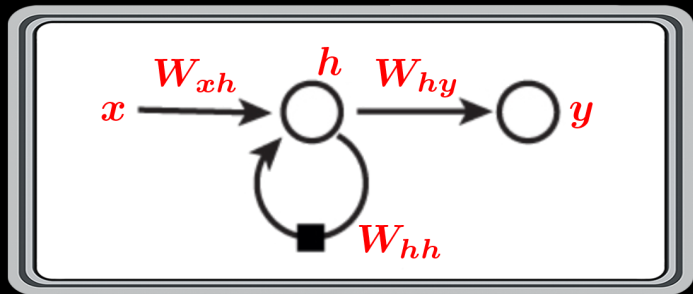
РЕКУРРЕТНАЯ СЕТЬ

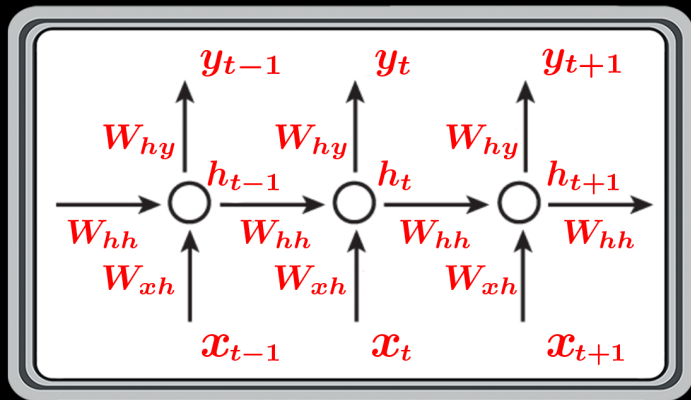
- Последовательно получает токены x на вход
- Обновляет скрытое состояние h
- Выдаёт ответ y



$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1})$$

$$y_t = g(W_{hy}h_t)$$





- › Обратное распространение ошибки на развёрнутой сети
- › Backpropagation Through Time (BPTT)

ПРИМЕР: ГЕНЕРАЦИЯ ТЕКСТА

- » Вход: слово
- » Выход: вектор вероятностей, число элементов равно размеру словаря (слова или символы)

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS)[<http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm> Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

ПРИМЕР: ГЕНЕРАЦИЯ ТЕКСТА

ПРИМЕР: ОБУЧЕНИЕ С УЧИТЕЛЕМ

- Генерация признаков с помощью нейросети:
 - ▶ Скрытое состояние после прохода по всему тексту
 - ▶ Агрегация скрытых состояний после каждого токена

- › Рекуррентные сети обрабатывают тексты последовательно
- › Информация о предыдущих токенах хранится в скрытом векторе
- › Можно генерировать тексты