МФТИ

word2vec

ПОХОЖИЕ СЛОВА

МФТИ

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

ПОХОЖИЕ СЛОВА

МФТИ

- «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

ВЕКТОРНЫЕ ПРЕДСТАВЛЕНИЯ СЛОВ

МФТИ

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

word2vec

МФТИ

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

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

word2vec

МФТИ

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

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

где вероятность вычисляется через soft-max:

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

СВОЙСТВА ПРЕДСТАВЛЕНИЙ

МФТИ

- 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: $\vec{\text{one}} - \vec{\text{uno}} + \vec{\text{four}} \approx \vec{\text{quarto}}$
- Среднее представление по всем словам в тексте — хорошее признаковое описание

word2vec И ОБУЧЕНИЕ С УЧИТЕЛЕМ

МФТИ

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

РЕЗЮМЕ

МФТИ

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